

ECONOMIC THEORY AND THE ANCIENT MEDITERRANEAN

DONALD W. JONES



WILEY Blackwell

VISIT...

LANZAROTE
Caliente.COM

Economic Theory and the Ancient Mediterranean

Economic Theory and the Ancient Mediterranean

Donald W. Jones

WILEY Blackwell

This edition first published 2014
© 2014 John Wiley & Sons, Inc.

Registered Office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices

350 Main Street, Malden, MA 02148-5020, USA

9600 Garsington Road, Oxford, OX4 2DQ, UK

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

The right of Donald W. Jones to be identified as the author of this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Jones, Donald W.

Economic theory and the ancient Mediterranean / Donald W. Jones.

pages cm

Includes index.

ISBN 978-1-118-62787-7 (cloth)

1. Mediterranean Region—Economic conditions. 2. Econometrics. 3. Mediterranean Region—History—To 476. I. Title.

HC31.J64 2014

330.937—dc23

2014001157

A catalogue record for this book is available from the British Library.

Cover image: Trading scene on bowl from Naucratis, 6th century BC. © Interfoto / SuperStock

Set in 9.5/11.5pt Minion by Laserwords Private Limited, Chennai, India

Contents

Preface	xiii
Acknowledgments	xvii
Introduction	1
<i>Rationale</i>	1
<i>Organization</i>	2
<i>Method</i>	3
<i>Reader Outcomes</i>	3
<i>Themes</i>	4
<i>Relevance and Applicability</i>	5
<i>References</i>	6
<i>Notes</i>	6
1 Production	8
1.1 <i>The Production Function</i>	9
1.2 <i>The “Law” of Variable Proportions</i>	11
1.3 <i>Substitution</i>	13
1.4 <i>Measuring Substitution</i>	15
1.5 <i>Specific “Functional Forms” for Production Functions</i>	16
1.6 <i>Attributing Products to Inputs: Distributing Income from Production</i>	17
1.7 <i>Efficiency and the Choice of How to Produce</i>	18
1.8 <i>Predictions of Production Theory 1: Input Price Changes</i>	20
1.9 <i>Predictions of Production Theory 2: Technological Changes</i>	21
1.10 <i>Stocks and Flows</i>	22
1.11 <i>The Distribution of Income</i>	23
1.12 <i>Production Functions in Achaemenid Babylonia</i>	25
<i>References</i>	26
<i>Suggested Readings</i>	27
<i>Notes</i>	27
2 Cost and Supply	29
2.1 <i>The Cost Function</i>	31
2.2 <i>Short Run and Long Run</i>	32
2.3 <i>The Relationship between Cost and Production</i>	33
2.4 <i>Producers’ Objectives</i>	34
2.5 <i>Supply Curves</i>	35

2.6	<i>Demands for Factors of Production</i>	40
2.7	<i>Factor Costs in General: Wages and Rents</i>	41
2.8	<i>Allocation of Factors across Activities</i>	43
2.9	<i>Organizing Production: The Firm</i>	43
2.10	<i>A More General Treatment of Cost Functions</i>	46
2.11	<i>The Economics of Mycenaean Vases, I: Supply and Cost</i>	47
2.12	<i>Accounting for Apparent Cost Changes in Minoan Pottery</i>	49
2.13	<i>Production in an Entire Economy: The Production Possibilities Frontier</i>	50
	<i>References</i>	52
	<i>Suggested Readings</i>	53
	<i>Notes</i>	53
3	Consumption	55
3.1	<i>Rationality of the Consumer</i>	57
3.2	<i>The Budget</i>	57
3.3	<i>Utility and Indifference Curves</i>	58
3.4	<i>Demand</i>	60
3.5	<i>Demand Elasticities</i>	63
3.6	<i>Aggregate Demand</i>	65
3.7	<i>Evaluating Changes in Wellbeing</i>	66
3.8	<i>Price and Consumption Indexes</i>	70
3.9	<i>Intertemporal Choice</i>	73
3.10	<i>Durable Goods and Discrete Choice</i>	75
3.11	<i>Variety and Differentiated Goods</i>	79
3.12	<i>Value of Time and Household Production</i>	82
3.13	<i>Risk, Risk Aversion, and Expected Utility</i>	86
3.14	<i>Irrational Behavior</i>	88
3.15	<i>Fixed Prices</i>	90
3.16	<i>Applying Demand Concepts: Relationships between Housing Consumption, Housing Prices, and Incomes in Pompeii</i>	93
3.17	<i>The Economics of Mycenaean Vases, II: Demand</i>	96
	<i>References</i>	99
	<i>Suggested Readings</i>	99
	<i>Notes</i>	100
4	Industry Structure and the Types of Competition	103
4.1	<i>Perfect Competition</i>	104
4.2	<i>Competitive Equilibrium</i>	106
4.3	<i>Monopoly</i>	108
4.4	<i>Oligopoly</i>	110
4.5	<i>Monopolistic Competition</i>	111
4.6	<i>Contestable Markets</i>	112
4.7	<i>Buyer's Power: Monopsony</i>	113
4.8	<i>The Economics of Mycenaean Vases, III: Industry Structure</i>	114
4.9	<i>Ancient Monopoly and Oligopoly: Religion and Foreign Trade</i>	115
	<i>References</i>	117
	<i>Suggested Readings</i>	118
	<i>Notes</i>	118
5	General Equilibrium	120
5.1	<i>General Equilibrium as a Fact and as a Model</i>	120
5.1.1	<i>The facts</i>	121
5.1.2	<i>The models</i>	121
5.1.3	<i>The questions</i>	123

5.2	<i>The Walrasian Model</i>	124
5.3	<i>Exchange</i>	127
5.4	<i>The Two-Sector Model</i>	128
5.4.1	The basics with the Lerner–Pearce diagram	128
5.4.2	Growth in factor supplies	130
5.4.3	Technical change	132
5.5	<i>Existence and Uniqueness of Equilibrium</i>	133
5.6	<i>Computable General Equilibrium Models</i>	134
	<i>References</i>	136
	<i>Suggested Readings</i>	137
	<i>Notes</i>	137
6	Public Economics	139
6.1	<i>Government in the Economy: Scope of Activities, Modern and Ancient</i>	139
6.2	<i>Private Goods, Public Goods, and Externalities</i>	141
6.2.1	Private goods	141
6.2.2	Public goods	142
6.2.3	Externalities	143
6.3	<i>Raising Revenue</i>	149
6.3.1	Taxation 1: rationales and instruments	149
6.3.2	Taxation 2: effects of taxes	154
6.3.3	Taxation 3: tax incidence (who really pays?)	165
6.3.4	Taxation 4: optimal tax systems	169
6.3.5	Other revenue sources	173
6.4	<i>The Theory of Second Best</i>	174
6.5	<i>Government Productive Activities</i>	175
6.5.1	Public production and pricing	175
6.5.2	The supply of public goods and social choice mechanisms	181
6.5.3	Public investment and cost–benefit analysis	186
6.6	<i>Regulation of Private Economic Activities</i>	191
6.6.1	Rent seeking	192
6.6.2	The costs of regulation: the Averch–Johnson effect	193
6.7	<i>The Behavior of Government and Government Agencies</i>	194
6.7.1	Theories of government	194
6.7.2	Theories of bureaucracy	195
6.7.3	Levels of government	196
6.8	<i>Suggestions for Using the Material of this Chapter</i>	196
	<i>References</i>	197
	<i>Suggested Readings</i>	199
	<i>Notes</i>	199
7	The Economics of Information and Risk	202
7.1	<i>Risk</i>	202
7.1.1	The ubiquity of risky decisions	203
7.1.2	Concepts and measurement	205
7.1.3	Risk and behavior: expected utility	209
7.1.4	Risk versus uncertainty: the substance of probabilities	215
7.2	<i>Information and Learning</i>	217
7.2.1	The structure of information	217
7.2.2	Learning as Bayesian updating	218
7.2.3	Experts and groups	223
7.3	<i>Dealing with Nature's Uncertainty</i>	225
7.3.1	Contingent markets	225
7.3.2	Portfolios and diversification	230

7.4	<i>Behavioral Uncertainty</i>	235
7.4.1	Asymmetric information: problems and solutions	236
7.4.2	Strategic behavior	242
7.5	<i>Expectations</i>	246
7.5.1	The role of expectations in resource-allocation decisions	247
7.5.2	Adaptive models of expectations	247
7.5.3	The rational expectations hypothesis	249
7.6	<i>Competitive Behavior under Uncertainty</i>	252
7.6.1	Production behavior	252
7.6.2	Search problems	253
7.7	<i>Suggestions for Using the Material of this Chapter</i>	253
	<i>References</i>	254
	<i>Suggested Readings</i>	255
	<i>Notes</i>	255
8	Capital	258
8.1	<i>The Substance and Concepts of Capital</i>	258
8.1.1	Capital as stuff	259
8.1.2	Capital in the production function	262
8.1.3	Stocks, flows, and accumulation	263
8.1.4	Prices and values	264
8.1.5	Temporal aspects of capital	265
8.1.6	Measuring capital	268
8.1.7	The labor theory of value	269
8.2	<i>Quasi-Rents</i>	270
8.3	<i>Interest Rates</i>	272
8.4	<i>The Theory of Capital</i>	276
8.4.1	Present and future consumption, investment, and capital accumulation	276
8.4.2	Demand for and supply of capital: flows and stocks	279
8.4.3	Capital richness and interest rates	283
8.5	<i>Use of Capital by Firms</i>	284
8.5.1	Investment	284
8.5.2	Maintenance	287
8.5.3	Scrapping and replacement	289
8.6	<i>Consumption and Saving</i>	290
8.6.1	Intertemporal utility maximization	290
8.6.2	Hypotheses about consumption	291
8.6.3	Individual and aggregate savings	294
8.7	<i>Capital Formation</i>	294
8.8	<i>Suggestions for Using the Material of this Chapter</i>	296
	<i>References</i>	297
	<i>Suggested Readings</i>	298
	<i>Notes</i>	298
9	Money and Banking	301
9.1	<i>The Services of Money</i>	302
9.1.1	Money as a medium of exchange	302
9.1.2	Money as a store of value	302
9.1.3	Money as a unit of account	303
9.1.4	Stability of value	303
9.1.5	Monetization prior to currency	303
9.2	<i>The Types of Money</i>	304
9.2.1	Commodity money	304
9.2.2	Credit money	304
9.2.3	One special case of credit money: bank money	305

9.3	<i>Some Preliminary Concepts</i>	305
9.3.1	The price level	305
9.3.2	Inflation	306
9.3.3	“Nominal” versus “real” distinctions	307
9.3.4	What people in antiquity knew	309
9.4	<i>The Demand for Money</i>	309
9.4.1	Measuring money	310
9.4.2	The distinctiveness of the demand for money	311
9.4.3	Monetary theory and macroeconomics for ancient economies?!	312
9.4.4	The neoclassical quantity theory	313
9.4.5	Keynesian monetary theory	315
9.4.6	The contemporary synthesis	317
9.5	<i>The Supply of Money</i>	318
9.5.1	Supply of a commodity money	320
9.5.2	Creation of money by banks	323
9.5.3	The banking firm	328
9.5.4	Financial intermediation	332
9.5.5	Exogeneity / endogeneity of money supply and foreign exchange	335
9.5.6	Seigniorage: making money by issuing money	336
9.5.7	Bimetallism	337
9.6	<i>Inflation</i>	337
9.6.1	Causes of inflation	338
9.6.2	Mechanisms of inflation	339
9.6.3	Consequences of inflation	340
9.7	<i>Monetary Policy</i>	342
9.7.1	The players and their motives	342
9.7.2	Choice of monetary standard	343
9.7.3	Influencing the supply of money	343
9.7.4	Influencing the demand for money	345
9.7.5	International monetary policies	345
9.8	<i>Suggestions for Using the Material of this Chapter</i>	345
	<i>References</i>	345
	<i>Suggested Readings</i>	347
	<i>Notes</i>	347
10	Labor	350
10.1	<i>Applying Contemporary Labor Models to Ancient Behavior and Institutions</i>	350
10.2	<i>Human Capital</i>	353
10.2.1	Investment in human capital	354
10.2.2	Health	356
10.2.3	Guilds, occupational licensing, and entry restriction	356
10.3	<i>Labor Supply</i>	357
10.3.1	Utility analysis of individual and family labor supply	357
10.3.2	Lifecycle / dynamic labor supply	364
10.3.3	Supply of labor to activities	368
10.3.4	Household production	369
10.4	<i>Labor Demand</i>	375
10.4.1	The productive enterprise’s demand for labor	376
10.4.2	Derived demand	379
10.5	<i>Labor Contracts</i>	384
10.5.1	Information problems and incentives	384
10.5.2	The basis of pay	385
10.5.3	Sequencing of pay	387
10.5.4	Compensating differentials in wages	387

10.6	<i>Migration</i>	391
10.6.1	Economic incentives for migration	392
10.6.2	Consequences of migration	394
10.6.3	Refugee migration	396
10.6.4	Equilibrating migration flows when the wage rate doesn't adjust	396
10.7	<i>Families</i>	398
10.7.1	Marriage	398
10.7.2	Intrafamily resource allocation	405
10.7.3	Children and the economics of fertility and child mortality	412
10.8	<i>Labor and the Family Enterprise</i>	414
10.8.1	The farm family household and the separability of production decisions from consumption decisions	415
10.8.2	Effects of missing markets on labor allocation	418
10.8.3	Restrictions on household activities	420
10.8.4	Implications of the family farm model	422
10.9	<i>Slavery</i>	423
10.9.1	The supply of slaves	424
10.9.2	The demand for slaves	426
10.9.3	Investment in slaves	427
10.9.4	Market consequences of slaves	427
10.9.5	Slaves' incentives	427
10.10	<i>Suggestions for Using the Material of this Chapter</i>	428
	<i>References</i>	429
	<i>Suggested Readings</i>	432
	<i>Notes</i>	433
11	<i>Land and Location</i>	440
11.1	<i>The Special Characteristics of Land</i>	440
11.2	<i>Land as a Factor of Production</i>	441
11.2.1	Supply	441
11.2.2	Demand	441
11.3	<i>The Location of Land Uses</i>	442
11.3.1	The Thünen model	442
11.3.2	The bid-rent function	447
11.3.3	Equilibrium in a region	450
11.3.4	Modifying the social context	451
11.4	<i>The Location of Production Facilities</i>	452
11.4.1	Individual facilities	452
11.4.2	Industries	455
11.5	<i>Consumption and the Location of Marketing</i>	457
11.5.1	The structure of transportation costs	457
11.5.2	The shopping tradeoff: frequency versus storage	458
11.5.3	Aggregate demand in a spatial market	460
11.5.4	Hierarchies of marketplaces: central place theory	461
11.5.5	Periodic markets	462
11.6	<i>Transportation</i>	463
11.6.1	Infrastructure	463
11.6.2	Equipment	465
11.6.3	Pricing of transportation services	465
11.7	<i>Suggestions for Using the Material of this Chapter</i>	467
	<i>References</i>	468
	<i>Suggested Readings</i>	469
	<i>Notes</i>	470

12	Cities	472
12.1	<i>Cities and their Analysis, Modern and Ancient</i>	472
12.1.1	Classifying cities	472
12.1.2	Characteristics of cities	473
12.1.3	What goes on in cities	473
12.1.4	Ancient observations and contemporary analytical emphases	474
12.2	<i>Economies of Cities</i>	475
12.2.1	Scale economies in production	475
12.2.2	Externalities	477
12.2.3	Types of production	477
12.3	<i>Housing</i>	479
12.3.1	The Special Characteristics of Housing	479
12.3.2	Housing supply	480
12.3.3	Housing demand	481
12.4	<i>Urban Spatial Structure</i>	482
12.4.1	The monocentric city model	483
12.4.2	Multiple categories of residents	488
12.4.3	Working at home	489
12.4.4	Endogenous centers	490
12.4.5	Density gradients and the ancient city	491
12.4.6	Wage differentials across cities	491
12.5	<i>Systems of Cities</i>	492
12.5.1	Production and consumption within any city	493
12.5.2	Different types of cities	497
12.5.3	The city size distribution and its responses to various changes	499
12.6	<i>Urban Finance</i>	503
12.6.1	Local public goods	504
12.6.2	What to supply and how much	505
12.6.3	Raising revenue	506
12.7	<i>Suggestions for Using the Material of this Chapter</i>	507
	<i>References</i>	508
	<i>Suggested Readings</i>	510
	<i>Notes</i>	511
13	Natural Resources	516
13.1	<i>Exhaustible Resources</i>	517
13.1.1	The theory of optimal depletion	517
13.1.2	Different deposits	520
13.1.3	Uncertainty	521
13.1.4	Exploration	521
13.1.5	Monopoly	523
13.2	<i>Renewable Resources</i>	524
13.2.1	Biological growth	524
13.2.2	Harvesting	525
13.2.3	The theory of optimal use	527
13.2.4	Open access and the fishery	528
13.3	<i>Resource Scarcity</i>	531
13.4	<i>The Ancient Mining-Forestry Complex</i>	531
13.5	<i>Suggestions for Using the Material of this Chapter</i>	532
	<i>References</i>	533
	<i>Suggested Readings</i>	533
	<i>Notes</i>	533

14	Growth	535
14.1	<i>Introduction</i>	535
14.1.1	Economic growth: delimiting the scope	535
14.1.2	Growth in antiquity: is there anything to explain?	536
14.2	<i>Essential Concepts</i>	536
14.2.1	Production functions again	536
14.2.2	Technical change	537
14.2.3	Growth versus development	537
14.3	<i>Neoclassical Growth Theory</i>	538
14.3.1	The Solow model	538
14.3.2	Technology and growth in the Solow model	541
14.3.3	Endogenizing technical change	543
14.3.4	Extent of the market, division of labor, and productivity	545
14.4	<i>Structural Change</i>	546
14.4.1	Sectoral concepts as organizing devices	546
14.4.2	A two-sector model of an economy	548
14.4.3	Some stylized facts	549
14.5	<i>Institutions</i>	551
14.5.1	Property rights	552
14.5.2	Governments	552
14.5.3	Stability and change	553
14.6	<i>Studying Economic Growth in Antiquity</i>	553
14.6.1	What there is to explain	554
14.6.2	Organizing inquiry about economic growth with the help of growth theory	554
14.6.3	Studying episodes of growth following declines: beyond growth theory	557
14.6.4	Summary	559
14.7	<i>Suggestions for Using the Material of this Chapter</i>	559
14.7.1	Evidence of growth	559
14.7.2	Sectoral structure	561
	<i>References</i>	561
	<i>Suggested Readings</i>	564
	<i>Notes</i>	564
	Index	569

Preface

The goal of this volume is to provide scholars of the ancient Mediterranean region with an additional set of intellectual tools to support their research. Interest in the economic lives of people and societies in antiquity is longstanding, and over the last several decades, scholars have addressed topics involving economic growth, locational advantage, national income accounting, banking and finance, to name a few, sometimes appealing to concepts from contemporary economics. Closer familiarity with a wider range of contemporary economic concepts that may be useful in specific instances cannot but help students of antiquity accomplish their primary purposes. Adding these tools to those already brought to bear from neighboring social science fields – anthropology, political science, sociology, linguistics – as well as tools from physical sciences and engineering, will add to the resources that can be brought to bear on research into life in antiquity.

There are many excellent introductory economics texts. While they are accessible to the general student who does not plan to study further economics, they also lay the foundations for the student who will go on to make contributions to economic science. This handbook is designed expressly for the student who has a demand for relatively advanced concepts in economics but whose goals are to make contributions to understanding the histories or prehistories of ancient societies in the Mediterranean and Aegean regions. Consequently, I have developed

a combination of basic concepts, presented compactly but intuitively, and more sophisticated concepts that will prove useful in applications to a wide range of social problems. The present volume provides pure theory, but with an emphasis on the practical applications of the models.

Economics is not a particularly easy subject, but then neither are ancient, inflected languages. The student of ancient languages might take some comfort from realizing that the conjugation of verbs and declension of nouns, pronouns and adjectives is essentially application of the calculus process of differentiation of the stems of those words. Reading the texts, which requires the student to infer the base word from the endings (not all of which may come at the end!) is equivalent to integrating a differentiated function back to its original form. The fact that linguists have been using computers to conduct analyses on various aspects of languages highlights the mathematical properties of the logic of languages.

Economics is a discipline without a whole lot of facts; it brings to the table primarily logic, with rules about how to apply the logic to empirical observations. However, if the logic does not apply to observable human behavior it is of little ultimate interest in a social science. Economists are proud to point to Nobel Prize-winning physicists who took up physics because economics was too hard, but in fact those famous physicists who switched to physics from economics because economics was too boring or too easy are just about as numerous as those who switched because

it was too hard. As Milton Friedman, a Nobel Prize-winning economist, has said frequently in classes, many concepts in economics can take quite a while to understand, but once you finally *do* understand them it's usually difficult to fathom how you ever *failed* to understand them. Most of it is just common sense. Another prominent economist recently opined that economics is harder than physics but easier than sociology, because of the degree to which issues "stand still" for analysis in the three subjects. Had he thought about it, he might have put the study of ancient societies to the harder side of sociology.

Contemporary economics is a thoroughly mathematized social science, possibly because so many of the phenomena to which it directs its attention lend themselves well to measurement. It is difficult to explain much of economic theory without using any mathematics at all – many of the introductory textbooks that avoid mathematics run 600 or 700 pages and even longer – just to introduce the very basics, and even then, frequently with mind-numbing tables of numbers to convey points that could be made much more compactly. Numerical examples are quite useful but I have largely avoided them in favor of diagrams and simple formulas which take the reader no further into mathematical science than the four basic arithmetic operations (addition, subtraction, multiplication, and division). These formulas can be read just like text: "the price times the quantity equals the amount paid (or received) ..." I am aware that many in the audience will have limited patience for plowing through reams of abstract material before they get to results they realize they can use in their own business of understanding ancient societies. I have tried to find a tradeoff between compactness of presentation and intuitive explanation that will permit these students to progress rapidly through the rich offerings of economics and take away concepts they can use immediately, without immersion in a three- or four-year, intensive program in economics.

I have avoided attempting to eschew all so-called jargon, which is simply the pejorative terminology for the technical lexicon that economists have developed to communicate

professionally. Most disciplines have developed one- or two-word terms for concepts that could take a paragraph or more to refer to otherwise, and economics is no exception. Since one of the goals of the volume is to prepare archaeologists, ancient historians, and philologists to enter the professional economics literature themselves according to their needs, they can spend several months to a year or more picking up the technical lexicon, with all its abbreviations, variants, and shorthands, on their own, or I can offer a quick and compact – and, I hope, "user-friendly" – introduction to it here. I thought the latter made more sense. A final word about the structure of the book. The first five chapters present the core of economic theory, and serve as textbook as much as handbook. The remaining nine chapters apply the basic principles of the first five chapters to present major results from substantive fields of economics such as taxation, labor, and so on. It will be difficult to get a lot out of these last nine, handbook-style chapters without understanding the first five: the formulaic notation could appear difficult, and the expositions use a number of sophisticated concepts that are developed intuitively in the first five chapters. If you see them for the first time in, say, Chapter 6, their use may seem unforgiving. However, if you read the five core chapters initially, you may feel like the young Mark Twain observing his father's growth: surprised how much you've learned in those chapters.

Some scholars may believe that the ancient world does not offer enough "data" to make investment in theory worthwhile. Any inferences made on the basis of observations use theory. If the observer-explainer is not aware of the theory he or she is bringing to the observations, there is little assurance that the implicit theory being used has the properties of logical coherence and compatibility with other sets of observations that the very observer-explainer would want a theory to have. An abundance of data makes at least some accounting framework obviously valuable; lots of observations can be made with only implicit theory before one begins to notice the weaknesses deriving from the lack of an explicit body of theory. Data-poor situations place scholars in

the position, very early in an investigation, of asking “What can these observations *mean*?” An explicit theoretical framework can offer valuable guidance immediately, helping to connect dots, as it were, and offering restrictions on possible explanations.

While I have billed this book as targeted at students of the ancient Mediterranean (liberally defined to include the Aegean, Black Sea,

Arabian/Persian Gulf, and Red Sea regions as well), scholars of antiquity in other regions – the Americas, east and south Asia, northern Europe, and so on, might find the theories useful, even if the examples fall outside their areas of interest. The regional emphasis is a reflection of my own background and experience rather than an implicit statement on the applicability of the theory.

Acknowledgments

Two debts I must acknowledge first: to Geraldine Gesell, who introduced me to Cretan archaeology and has given me some two-and-a-half decades of sage advice and unstinting encouragement; and to the late Elizabeth Lyding Will, who, in addition to much other good advice and support, first suggested writing down lecture notes on economics for an archaeological audience and made a number of specific suggestions on early versions of chapters, which I have endeavored to incorporate in this manuscript.

A number of other people have offered useful suggestions, ideas, assistance of varying sorts, and encouragement over the course of the preparation of this manuscript and beyond. Alphabetical is the best order in which to acknowledge them: Henry Colburn, Alexander Conison, Michael Leese, Susan Martin, Charlotte Maxwell-Jones, William Parkinson, David Tandy, Aleydis van de Moortel, and David Warburton.

I offer my thanks to the Department of Classics of the University of Tennessee for maintaining me as an adjunct professor over the past decade-and-a-half, with the library access that has offered.

Two readers for Wiley Blackwell offered encouragement and useful suggestions, and one of them, who read the entire manuscript, subsequently made a number of very helpful suggestions, for which I am quite grateful. And I am grateful to my editors, Haze Humbert, Allison Kostka, and Ben Thatcher, and others of Wiley Blackwell for their help. Ashley McPhee, Wiley Blackwell's Editorial Assistant for Classics and Ancient History, and her cover designer, Yvonne Kok, developed a handsome array of cover design choices. And finally, I appreciate the splendid work of the freelance editorial and indexing team, led by Nik Prowse, the project manager; and including David Michael, copy-editor; Felicity Watts, proofreader; and Neil Manley, indexer.

Introduction

This volume's primary goal is to offer a compact, if intense, introduction to contemporary economic theory for scholars of the ancient Mediterranean-Aegean-Near Eastern region: archaeologists, ancient historians, and philologists.¹ Why might people from these fields find this material of interest? Economic topics of antiquity have been of abiding interest in these related disciplines, and while much has been learned over the past four decades or so, that scholarship has, frankly, been hampered by misconceptions about and awkward applications of the body of theory that offers direct insights into those topics. It takes long and difficult effort to turn oneself into a classicist or a scholar of the ancient Near Eastern or Egyptian languages, plus acquire the modern languages necessary to read the present literature in the field, plus learn the history, archaeology, methodologies, including some physical science applications, and the list could go on. There's a lot to learn, and limited time and energy. Triage principles certainly have to be applied. That, and, we might as well be direct about it, a lot of twentieth-century political and ideological baggage has attached itself to economic theory in a number of social science and humanities disciplines.

Rationale

It is not difficult to find comments in the literature on ancient societies and economies that, when stripped of their costumery, amount to "economic theory isn't applicable." Reading some of these works, it is not clear that the knowledge base is always sufficient to reach such a conclusion. Contemporary economic theory can model the maximization of prestige as comfortably as it can profit, and there is no reason to view the theory as a reductionist tool.² Other scholars of antiquity find economics and economic models useful but sometimes their use of them is hampered by limited understanding of how the models work. Nobody has enough time to study everything that really needs to be studied, and learning enough economics to be functional with its concepts takes time and energy but, at some point, excuses involving time and energy become threadbare. This volume offers humble assistance in surmounting these twin problems of time limitations and unfamiliarity. At the very least, the volume offers some archaeologists, ancient historians, and philologists the opportunity to make more authoritative statements of "economics isn't applicable," and explain why they reach such a

conclusion, from a base of sound knowledge; some may find a modest conversion with closer understanding; and others who found the models appealing all along may find them more useful.

In trying to make the essentials of economic theory and a broad swath of its principal applications more accessible to busy scholars in other fields, I have not been able to make the subject easier than it is. I regret that, but then no one has been able to make Attic Greek or Middle Egyptian easy either. Plenty of late-night tears have been shed during the youth of all these disciplines' current elders. However, I hope I have succeeded in some measure in making the operation a quicker affair than it otherwise would be: a few weeks or months of pain versus several years of classes. The introductory and intermediate textbooks dealing with the topics of this volume's chapters easily would stand five or six feet high. In making this distillation, I have omitted the numerical examples, case studies, and problem sets that inflate the page counts of these textbooks but undoubtedly assist the novice's learning. I have, however, offered examples of cases from antiquity to demonstrate how the theories might be applied to subjects of interest to the target audience. This approach assumes that the audience consists of scholars of a certain degree of personal and intellectual maturity and intellectual discipline, who understand from experience the effort required to surmount entry barriers to new fields of study, be they another ancient or modern language or even a new, relatively unexplored topic of inquiry. These scholars have acquired the patience to read a sentence several times, if necessary, to understand it. They are accustomed to flipping back a few pages in a text now and then to re-establish a continuity of thought if necessary.

Organization

The organization of the volume is as follows. The first five chapters form the core of economic theory – called price theory or microeconomics. The following nine chapters treat major applied branches of economics. They are applications of the basic price theory addressed in the first five chapters. However, many of these chapters introduce additional models of issues that involve basic price theory. Two examples: the theory

of externality is treated in Chapter 6 on public economics, and risk is the subject of Chapter 7. The externality concept³ requires the concepts of private and public goods, which need not burden the reader just struggling with the concepts of supply and demand, who doesn't really need to know at that point that goods that are supplied and demanded may have sharply differing – or blurred on occasion – economic properties: better to stick with simple, privately consumed goods such as food and clothing before moving on to bridges and city walls, which are consumed by many people at the same time. In the case of risk, we all know that it's ubiquitous but, again, the consequences of risk are better appreciated when they can be compared with situations without risk. Again, it's a needless complication on the first date. There is reason to the order of appearance of the applications chapters. Each introduces some new concepts beyond those treated in the core chapters – but using the principles of the core chapters. Later applied chapters often make recourse to principles developed in earlier applied chapters.

To attempt to illustrate how the models and reasoning introduced in these chapters can be applied to problems of interest to scholars of antiquity, the core chapters offer some sample application sections, and each of the nine applied chapters ends with a section containing suggestions for how the material of the chapter might be used by archaeologists or ancient historians and philologists (I group the latter two together on the grounds that both rely much more on textual evidence while the archaeologists rely more heavily on material evidence). These suggestions should not be considered definitive but rather as offering a few ideas, which may suggest applications to topics I have not thought of. Also, following the references to each chapter is a short list of suggested readings in economics pertinent to that chapter. Some of these works may strike readers as dated, with publication dates in the 1980s and even the 1970s, and earlier. The reasoning for my choices is several. First, some are simply classics and remain the best and most accessible statements on their subjects. Second, I have used these editions myself and understand them but have not actively pursued subsequent editions. While advances have been made in all these fields, the basic ideas to which readers are directed in

these works remain foundational, while some of the advances are simply less accessible. Third, in some cases of works that have gone through many editions, some still continuing, I frankly think an earlier edition is superior to later or even current ones for the purposes and needs of this volume's readers. And fourth, if a reader wants to buy some of these works, the older ones are generally considerably cheaper than newer ones.

Method

A note on what might be called method: when necessary, I have used symbolic expressions in the text. There's no way around the fact that these are mathematical expressions, or sentences, and presented to an audience many of whose members chose other routes in college. However, I have kept the mathematical operations they present to the four basic arithmetic operations – addition, subtraction, multiplication, and division⁴ – and these expressions can be read just like sentences: “This times that, plus the other thing times something else, equals what we're interested in.” These expressions can be read for precise meaning just as an ordinary sentence can be read – hence my inclusion of them in sentences rather than set apart from the text. It's important for readers to see exactly what is and is not included in an economic calculation. The application of economic theory can live with the imprecision frequently found in ancient textual and material evidence, but the logic of the theory requires fairly sharp dividing lines between what is and what isn't included in a concept, and the mathematical expression facilitates this precision. That said, I grant that some pretty fuzzy concepts can be wrapped up in a symbol to make them look crisper than they are, and readers must beware of such ornamentation. And finally, if a reader wants to get something out of articles in professional economics journals, the experience of settling down to read a mathematical expression in this text should open up more than the abstract and the conclusion (if those) in the typical technical paper.

A reader may be tempted to peek ahead at some chapters on particularly interesting subjects. For example, growth in antiquity (Chapter 14)

recently has become a topic of considerable interest to scholars of antiquity. A look-ahead is likely to be disappointing to readers who haven't absorbed at least a fair amount of the core five chapters' material. Each of the applied chapters (6–14) builds on the basic theory of the first five chapters, and some of the applied chapters introduce material that is used in subsequent chapters. For example, the economics of labor is the unrelenting application of production theory and demand theory to problems involving the relationships between households and the outside world, at any particular date in time and over lifetimes, even generations. Parts of it get complicated quickly even for economists accustomed to the models. Surely classicists and other philologists can think of texts they would recommend a student tackling only after working through a number of previous texts.

Reader Outcomes

In this volume, I hope to communicate how the logic of contemporary economic theory works, how the theory is applied to address particular questions, and how it can be applied to research topics in the economies of ancient Mediterranean societies. If I am successful, what are the reader outcomes I hope to achieve? First, a dedicated reader should emerge from the volume with a reasonably nuanced knowledge of contemporary economic theory and an understanding of some basics of its application. I recall a statement in a textbook early in my own economic education to the effect that most of the useful economics acquired during a Ph.D. program was, in effect, the basic principles learned in the sophomore course, simply applied in more intricate ways. There is, of course, more to it, but there is a large lump of truth to the statement. These basic principles can be learned efficiently, although learning to trust oneself with them, and spotting when and how to apply them, takes practice. Some economist a number of years ago said that the principal reason to learn economics was to be able to expose the nutty arguments of other economists, a statement with which many economists, particularly macroeconomists, probably would agree. This ability would be a useful reader outcome.

Second, the reader should acquire the lexicon of contemporary economics, which uses a combination of single words – nouns, adjectives, and verbs – that can pack a paragraph or so of meaning, and words that offer the potential confusion of having parallel common use as nontechnical terms in lay speech and different meanings in disciplinary technical applications, “demand” being the example *par excellence*, with “hedonic” not far behind. Familiarity with the terminology can help make scholarly expression more precise and effective as well as improve understanding of contemporary economic literature.

Third, the reader will emerge at the end of this tunnel with an appreciation of the architecture of contemporary economics – the names and subject matters of its various branches and fields of application. When searching for economic literature that may offer assistance in studying a problem in antiquity, the reader will have a better idea of where to look and more cogent key words to use. This said, I have been selective in my choices of topics for inclusion here, ignoring some interesting areas of theory for which I thought applications to problems of antiquity were simply too remote to justify burdening the reader further. Of course, these are judgment calls; some people may think I have included too many such topics anyway (clearly, I disagree though); others may think I have omitted some topics that warranted inclusion. For the latter category, readers who have made it through this volume should be able to use their knowledge of economics’ architecture to find those other theories.

Fourth, the reader will understand that there is no such thing as the “model of the Mycenaean palatial economy,” or the “model of the ancient Mesopotamian temple economy,” or the “model of the Roman economy,” or the “model of the [substitute time and place] economy.” Stated alternatively, there are so many models of the Mycenaean palatial economy, the Roman economy, *et al.*, as to make the term meaningless. There are as many models of specific issues in any of these times and places as scholars make to help themselves study their questions. Someone may assemble a general equilibrium model of the economy of some Mycenaean city state and see how its predictions accord with what is known or suspected from various records – and what the model predicts about things we can’t see today because all the evidence has rotted

away. Some scholar’s question might be much more restricted, not to say focused, than an entire economy, dealing with, say, why rural dwellers migrated to a major city such as Rome at a particular time. Addressing that question involves setting up a model, whether the model is explicitly mathematical or verbal, or whether the person asking the question even recognizes his thoughts as a model. One outcome of this text will be to encourage special-purpose model building as a way of thinking about economic issues in antiquity and to provide the tools with which to do so, again, whether a scholar desires to develop a mathematical, verbal or box-and-arrow-diagrammatic model.

Finally, for a special category of reader, a scholarly improvement would be a more informed basis for an attitude: a movement from, say, inherited prejudice to more reasoned and factual bases for sanction.

Themes

Several overarching themes recur throughout the chapters. In fact, it could be said that economics is the repeated application of them. First is the importance of incentives to individual behavior. Without being automatons, people respond to the incentives they face, although their responses may vary in strength. Some incentives derive from nature; others are socially or culturally created. In the longer term, some preferences may even be in part responses to incentives.

Second is the importance of what could be called the adding-up condition: when all the calculations are completed in the analysis of some problem, everything must be accounted for. Nothing is created or destroyed – everything comes from somewhere and goes somewhere. One incarnation of this principle is the budget constraint, whether applied to individuals or to firms or an entire economy or society: the availability of resources limits the range of actions and their outcomes.

Third is the limited importance of rationality. Commonly used as grounds for dismissing the applicability of economic theory to particular times and places, irrationality can be – in fact must be, if the term is to have any meaning – given precise definition. I consider the term “bounded rationality” a redundancy, since

it would be irrational to apply reason beyond the point where it stops helping. Rationality is sometimes conflated with knowledge, scientific or otherwise, but the two are entirely different. The assumption of rationality can be supplanted with alternative assumptions about behavior, with some predictions of rationality-based models emerging intact, others being modified.

Fourth is the importance of substitution: there is usually more than one way to skin a cat. People can accomplish the same goals in different ways and often do. Fifth is choice. People commonly, maybe even generally, face alternatives, and they decide which ones to avail themselves of.

Sixth is specificity, which must answer the question, “How does such-and-such happen?” When we propose the idea that, say, an ancient government caused something to happen or saw that its people did certain things, exactly how did it accomplish that action? Specificity provides a good laugh test for a lot of hypotheses.

This is a somewhat personal list, although I doubt most economists would raise major objections. Some might reorder them, add some items, or consider the overlap between some of the themes grounds for merger of some, but I think this list would find broad acceptance. These themes appear in the theories of other social science disciplines, so while economics has particular methods of implementing them, they are not disciplinarily exclusive concepts. In fact, much of economics at work on problems involves the application and interaction of these themes.

Relevance and Applicability

I brought up the topic of applicability, or relevance, early in this introduction, and it is worth getting more explicit about some of those concerns. The perspective of this volume is that contemporary economics offers very general models of resource allocation that can be tailored to many institutional settings. The basic price theoretic models of the behavior of individual agents impose no explicit institutional structure other than something that would let people find one another and guarantee they wouldn't be robbed or killed instead of traded with. In fact, introductory textbooks commonly appeal to Robinson Crusoe – pre-Friday – as a paradigm for understanding some of the basic principles

of economizing production and consumption behavior; the arrival of Friday just introduces exchange. Agricultural household models, which are introduced in Chapter 10 (labor), describe resource allocation by households, which only at a stretch could be considered formal markets themselves, with or without participation in exchanges outside the household. The scope for their application to problems of antiquity is very broad.

The concepts of prices and markets have accumulated some misperceptions; three in particular I will note here: that markets are required for prices and markets didn't exist (at least over much time and space); that prices can't exist without being denominated in currencies, which of course didn't exist in 2000 B.C.E.; and, when prices are accepted as having existed, that they were fixed by custom, possibly for long but indefinite periods.

Three points on the first issue. First, markets vary widely in organization, and they certainly don't require fixed stalls, auctioneers or whatever paraphernalia some scholars might want to attach to them. Second, they aren't necessary for exchange, which is widely reflected in implements found in Neolithic house remains that certainly weren't made by the occupants of each house. And third, the institutions surrounding exchanges surely evolved over time to accommodate more regular exchanges as specialization developed.

On the second issue, in any of these settings, a price is nothing more than the ratio at which two goods (or services) exchange for one another. It requires no currency and can be denominated in any good desired – which appears to be roughly how the Egyptian deben worked. People in antiquity surely had as good an idea how much something was worth to them as we do today and equally surely measured value in some metric that would let them make the comparisons they needed to make, even when the metrics weren't written down.

On the third issue, a useful way of thinking about long-term fixed prices may be to focus on the fact that these economies were predominantly agricultural – probably 90% or more of what today we would call their gross domestic product consisted of agricultural products. Agriculture is notoriously subject to weather variations, and the Mediterranean region includes

many areas of considerable annual variability. That means that, in some years, some crops came up a lot shorter than other crops and were much scarcer relative to nonagricultural goods, as well as less affected agricultural goods, than they were in other years. People in rural communities surely exchanged things with one another routinely; if it was customary to keep, say, wheat at an unchanging ratio relative to, say, wood chips, and there was a particularly bad year for wheat that left the availability of wood chips unchanged, do we really think that some fortunate person with a large accumulation of wood chips would have eaten quite well in one of these bad years by demanding the exchange of wheat for wood chips at the customary ratio, while his unluckier neighbors would have starved but with a lot of wood chips to burn when they ran out of wheat at the customary ratio? Expressed so baldly, this result is absurd, and no scholar working in the customary-price modeling tradition would ascribe to it. To preserve the model of unchanging prices from such an absurd outcome, we could add a side condition – say, along the lines of it also having been customary for no one to make such a call upon the neighbors during such a bad season – which recognizes that the first custom doesn't work under certain conditions, and could itself require some further side conditions for logical reconciliation. The model gets more and more complicated as we pile caveats onto the alleged customs until the structure caves in on itself of its own weight. Surely exchange ratios changed to reflect relative scarcities, even when they have not left records – although Babylonian records most certainly do demonstrate frequently changing relative prices of agricultural goods.

A final topic to address in this introduction also concerns the issue of the relevance of economic theory, but from the perspective of model assumptions. As a simple example, consider the Thünen model of land use around a central collection point, a model that has become reasonably well known among archaeologists and ancient historians interested in spatial organization of activities. Every introductory exposition of the Thünen model begins by clearing out the model landscape of all features that would interrupt transportation, yielding the infamous transportation surface, a patently unrealistic landscape. To reject the model as inapplicable or irrelevant to, say, the Apennine region of Italy or the central Peloponnese because of the mountains misses the point that the most stringent assumptions of the model simply provide a baseline against which to evaluate how departures from it would affect its predictions about the effects of transportation costs on what people do at and between different locations. It's harder to haul stuff in hilly or mountainous areas so transportation costs are higher there, and the model tells what happens when transportation costs are higher, not necessarily everywhere, but even just somewhere. Replace the flat-plain assumption, turn the crank on the model, and look at the more tailored result. Assumptions can be changed, and these models, themselves being things that are made, can be subjected to major structural modifications to accommodate specific circumstances. As teaching devices, the simpler models are more effective. Using parts of various theories to build a model of a specific issue involves a good bit of learned art as well as science.

References

-
- Dickinson, Oliver. 2006. *The Aegean from Bronze Age to Iron Age: Continuity and Change between the Twelfth and Eighth Centuries BC*. New York: Routledge.
- Jones, Donald W. 1999. "The Archaeology and Economy of Homeric Gift Exchange." *Opuscula Atheniensia* 24: 9–24.

Notes

- 1 I offer this regional restriction only to reflect my own knowledge base and the examples I use in the text. The subject matter is equally applicable to study of, say, Chinese or South American archaeology and ancient history as to that of the Greeks, Romans, Egyptians, and Mesopotamians.

- 2 I have striven to do this myself in Jones (1999, 23) especially regarding the specific way prestige is maximized. A prominent Mycenologist has recently referred to this article as neglecting traded goods (Dickinson 2006, 206), a correct observation that could be modified by adding one term to expression A.2, two terms to expression A.3 (Jones 1999, 20), and two additional relationships characterizing the acquisition process, adding two endogenous variables and increasing the six-equation system of expression A.20 (Jones 1999, 22) to an eight-equation system. A six-equation system has potentially 36 terms; an eight-equation system potentially 64. Some of the terms in each system will be zero, as not all variables interact with all others, but there is a cost to additional information about trade in terms of understandability of results.
- 3 If trade is considered to be a sufficiently important part of the problem, further study of the entire representation might find other simplifications that could reduce the cost of adding trade. This is an example of why economists try to keep their models as parsimonious as possible, which can appear to people outside the field of economics as unreasonably unsatisfying.
- 4 Basically, various forms of bothering your neighbor, from keeping him awake at night with your parties to dropping soot from your chimney onto his clean laundry to polluting his stretch of the creek with your sheep.
- 4 For the record, I keep the more intricate calculations out of sight and just report results that are expressible with the four arithmetic operations.

Production

Production is possibly the basic economic activity. Without it there would be nothing to consume, so the theory of demand would not be much of an issue. Consequently we begin our introduction to contemporary economic concepts with the choices people face when producing goods or services. In addition to introducing you to a particular body of theory, we also begin here in exposing you – gradually though – to the terminology of contemporary economics. Much of it is intuitive, but at just enough of an oblique angle to daily meanings of the identical words that you should pay careful attention. Our beginning point is the relationship between the things people use to produce other things and the things they produce with them – called inputs and outputs in the economic lexicon. The concept of the production function (sections 1.1 and 1.2) makes aspects of these relationships somewhat more precise than their use in casual conversation, but the degree of precision can vary according to the need for precision, which is a pleasant characteristic of this body of theory. The production function characterizes the technology – the actual physical and engineering relationships among inputs and outputs – in a fashion that constrains the choices people find it useful to make as well as the consequences of any choices they do make.

Correspondingly, changes in technology can change both choices and results (section 1.9).

One of the more important insights that contemporary economics uses, time and again, is that there is generally more than one way to do just about anything. Economics calls this aspect of life “substitution” or “substitutability” (sections 1.3 and 1.4). It characterizes consumption as well as production, but in this chapter we’ll focus on its role in production choices. One of the critical capacities of contemporary production concepts in economics is the ability to attribute proportions of products to the inputs that helped produce them. This attribution is called income distribution, and it involves attributing the product(s) produced to the inputs that produced them (or their owners, more precisely) in the form of income (section 1.6). This process may actually feel quite intuitive to scholars of the ancient world who are accustomed to thinking of many workers, particularly in the Near Eastern and Aegean palatial and temple economies, being paid in the form of rations or a comparable part of what they produced. It’s the same thing, basically. (As an historical accident of intellectual development, the term “income distribution” has also come to name a different, but certainly not unrelated, concept – that of how a total

income in an economy is distributed among its members. This has become called the “personal distribution of income” to distinguish it from the “functional distribution of income,” which refers to how output is attributed, if not necessarily actually distributed, to the inputs that produced it; section 1.11.)

Throughout this introduction to concepts about the economics of production – the choices people make in production – we have woven both actual and hypothetical examples from times and places in the ancient Mediterranean region. We close the chapter with a more extended example of how the use of concepts from production theory can illuminate the interpretation, and possibly even the translation, of ancient texts.

Economic concepts are prescriptive, as well as descriptive, in the sense that they identify the choices people could make that would make them the best off, in their own assessments, in terms of their own goals. Accordingly, the concept of efficiency emerges (section 1.7). With the further step of a widespread belief that most people at most times and places haven’t willingly left “food on the table,” these descriptive prescriptions also yield predictions of how people will behave – the choices they’ll make – in a wide range of circumstances (sections 1.8 and 1.9).

1.1 The Production Function

The workhorse concept of the theory of production is the production function, which relates the quantity of a product produced to the quantities of things used to produce it. The “things used to produce it” are called “factors of production” (sometimes “factors” for short) or “inputs.” For expositional purposes it is common (because it is simple) to study production functions with two inputs. Suppose we consider cotton (an output) to be produced with labor and land as the inputs, or the factors of production. Introducing some simple notation, we could use the shorthand $Q = f(L, N)$, where Q represents the quantity of cotton produced, L is the quantity of land used, N is the quantity of labor used, and f stands for the technological relationship between the inputs and the output.¹ The expression $Q = f(L, N)$ is read as “ Q equals (or “is”) a function of L and N ,” not “ Q equals f times L or N .”

Assume that all units of labor are equivalent to one another (that is, no big strong fellows and small weak fellows), all units of land are identical (fertility, slope, and so forth), and that all units of the cotton are of the same kind and quality. Otherwise, how could we compare units with one another? If you wanted to distinguish between, say, two categories of labor, one small and weak, the other big and strong, you would just specify two different labor inputs. This is the first example of a simplifying assumption in economic analysis (most assumptions do simplify; life is complicated enough without assuming that it is more so). The second example is in the assumption that the production function has just two inputs in it. This is a commonly used assumption designed to highlight the behavior of an individual factor. We could have called one of the factors “labor” and the other “all other inputs.” A two-factor designation serves to demonstrate most – but admittedly not all – of the behavior we want to investigate in production. The same simplification to just two items will appear commonly throughout this survey.

The relationship between each input and the output is precisely defined. To get more cotton, if the quantity of land is held fixed at the amount $\bar{L}$, we must increase the quantity of labor used. Conversely, if labor is fixed at $\bar{N}$, to get more cotton we must increase the amount of land we use. To get more output, at least one of the inputs must be increased in number. Further, production functions commonly – but not necessarily always – have the property that if the quantity of any one of the inputs used (we are not restricted to only two inputs; this is just for expositional convenience) is zero, the output is zero. Thus, $Q = f(0, N) = f(L, 0) = f(0, 0) = 0$.

Production functions contain considerably more information about the technology of production than just that more inputs are required to produce more of any output. They describe (i) exactly how much more of each input is required to produce another unit of output, and how this quantitative relationship can be expected to change as quantities of inputs and production change; (ii) the ways that other inputs affect the relationship between any particular input and output; (iii) relationships among inputs such as substitutability and complementarity; and (iv) the effects, if any, of overall scale of production

on the productivity of inputs. They help predict the employment decisions of producers and how producers will respond to cost changes and various technological changes.

Even if ancient data are scarce or missing altogether, the concept of the production function is useful, simply for collecting and clarifying your thoughts about what was used in production and what factors might have caused production to differ among locations or times. When we want to use the production function concept to think about a particular line of production at a particular time and place, there is absolutely no difficulty in adding more factors of production than the two we've talked about so far. To think about the economics of, say, pottery production, we certainly would want to include labor time, and for a relatively large potting operation, possibly several skill levels of labor. On the other hand, we might decide that land used in pottery production is so insignificant that we could just ignore it; or alternatively, we might have a case of ceramic production in a city such as fifth-century Athens, where finding space to let freshly turned pots dry before firing, as well as space for kilns and fuel inventories, would have been a non-negligible concern. Next, we might have some capital equipment – wheels, brushes, various tools for smoothing and scraping. Then there is the clay itself, which may be quite specialized. The kilns for firing the pots are a type of capital equipment, and the fuel for the fire is a material input. Each of these inputs would have required decisions that the remainder of the chapter will examine: how much to use, proportions relative to one another, technically possible and economically (even aesthetically) acceptable substitutions among one another.

The pottery example is a case of a production function for a product. We can develop production functions for processes as well, such as different types of industrial heat generation (for ceramics, metallurgy, baking, and preparation of various materials) and chemical processes such as dyeing and oil purification. Some of these production functions could be thought of as nested, in the sense that many of the chemical processes require controlled heat as well as other inputs combined with the heat. Economics has

developed the “engineering production function,” which uses chemical, mechanical, and other engineering knowledge to develop empirical relationships between “economic” inputs such as quantities of materials and sizes (capacities) of capital equipment and quantities of these process outputs, such as the magnitude of processed oil, dyed textiles, or quantity of heat output (Chenery 1948; Smith 1961, Chapter 2; Marsden *et al.* 1974). Much of the literature on ancient technologies that addresses such topics as the techniques of firing pottery and related ceramic materials such as faience and glass, smelting metals, and the production and use of various chemicals such as cosmetics and dyes, focuses on the material components of recipes, frequently on steps in processes, and occasionally on firing temperatures.² Much of the recent, physical science analysis of metals and ceramics is essentially reverse engineering from slags in the case of metals and the actual pots in the ceramic cases, to infer firing temperatures and technological innovations in materials that permitted desired transformations to occur at lower temperatures.³ While considerable technological knowledge has derived from these investigations, they tend to yield impressions of (i) unique methods used at particular places and times, with deviations representing errors and (ii) different technologies in use to produce similar or identical products at different locations or times. The element of choice of technique within a given technology, which was capable of alternative implementations, gets downplayed in these approaches. This is not a criticism *per se*, since each analytical methodology offers a certain range of insights; overcoming such restrictions presumably is the motivation for continual calls for interdisciplinary analysis of the ancient world.

Smith's example of “multiple-pass regeneration processes” illustrates the types of choices emphasized by the production function construct (Smith 1961, 42–44). In this type of process, a mixture of reactants, such as a vegetable oil, is passed over a bed composed of some catalytic substance such as fuller's earth. The filtering operation saturates the clay adsorbent but it can be regenerated by washing and burning in a furnace, although the clay's adsorbing capacity

falls with each regeneration. Eventually, after a number of these regenerations, the adsorbent declines sufficiently in efficiency that it pays to begin operations with a new adsorbent charge. Smith uses the chemical engineering parameters relating number of passes and subsequent regenerations to adsorbent capacity, then, through a series of substitutions involving quantities of adsorbent (clay) and equipment capacity, derives a production function that says that for a given capacity of filtering equipment, the adsorbent input to the process per year can be reduced only by increasing the number of passes per cycle, which entails using the clay at a lower level of efficiency. A given quantity of filtered vegetable oil can be produced in a year with alternative combinations of equipment capacity and throughput of fuller's earth. This example speaks to findings of alternative material recipes and process steps in ancient industries. There is no necessary implication of different technologies; archaeologists may be observing different choices of production techniques within a given technology. Why they might make those different choices is the subject of section 1.7.

In the meantime, before leaving this introduction to the production function, let's listen to Moorey (1994, 144) on the variability in the ancient use of kilns:

Pottery kilns were always adapted to the peculiar circumstances of the situation, the resources available, and the type of pottery to be produced . . . Throughout, into modern times, "open" and "kiln" pottery firing, in single- or double-chamber structures, might be found side by side in the same workshop or settlement for the production of different types of vessels or various ceramic fabrics.

Moorey's first observation focuses on the choices available to the ancient potters in choosing the combination of capital and other inputs (primarily fuel, probably, but possibly clay as well). The second observation may be a case of either coexistence of different technologies or simply of different ratios of capital to other inputs within a single technology, with the choice of that ratio depending on clay quality (which we could

translate into alternative inputs) or even specific products to be produced, with the input ratio possibly influenced by the relative prices different fabrics or vessel types could command. This last interpretation takes us beyond the concepts we've introduced so far, so with this we return to the development of production theory.

1.2 The "Law" of Variable Proportions

Consider the issue of how output changes with changes in the quantities of inputs applied. Figure 1.1 shows how total output increases as the quantity of labor (N) increases, with the quantity of land (L) fixed. As drawn, the total product (the curve labeled TP) increases moderately at first, then increases more steeply, then has its increase begin to slow down, eventually go to zero, and finally turn down. In Figure 1.2, consider that we

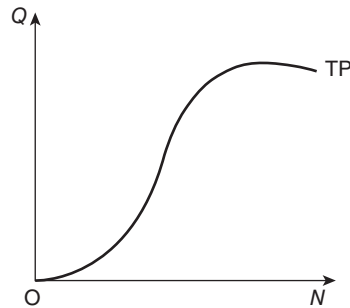


Figure 1.1 The total product curve.

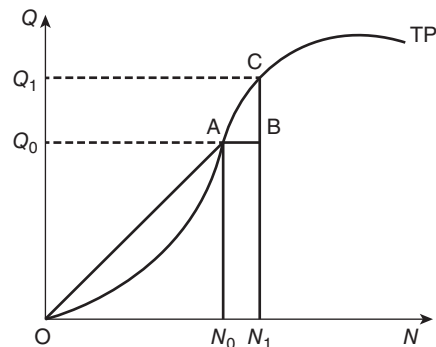


Figure 1.2 Average and marginal products.

have employed labor in the amount N_0 . The average product of labor (output Q_0 divided by labor N_0) can be represented by the slope line from the origin to point A on the TP curve ($Q_0 \div N_0$, or Q_0/N_0). Now, suppose we increase labor from N_0 to N_1 . Output increases from Q_0 to Q_1 , or to point C on the TP curve. The incremental output attributable to the incremental labor input is distance BC. This incremental output is called the *marginal product* of labor. (The definition of the marginal product of labor is $\Delta Q/\Delta N$, where the symbol Δ represents a change in the variable following it.) TP has some degree of curvature between points A and C, so we cannot draw any straight line to represent the marginal product. But suppose we contemplate making the difference between N_1 and N_0 smaller and smaller, until N_1 is just a tiny bit larger than N_0 – so close together that it looks like we are at a single point on the TP curve. The slope of the TP curve at point A (actually not a point, but the infinitesimal distance between N_0 and N_1 as we've shrunk the increment so much that we can approximate the difference by the point A) represents the marginal product of N at the quantity of labor N_0 . (The marginal product of labor at N_1 would be the slope of the TP curve at point C.)

The steepest line from the origin to a point on the TP curve will indicate the quantity of N per unit of Q (actually the Q/N ratio, which is the average product) that gives the largest average product of N . Figure 1.3 shows this line. The slope of this line equals the slope of the tangent to the TP curve at this point. So, at the maximum value of average product, average product (AP)

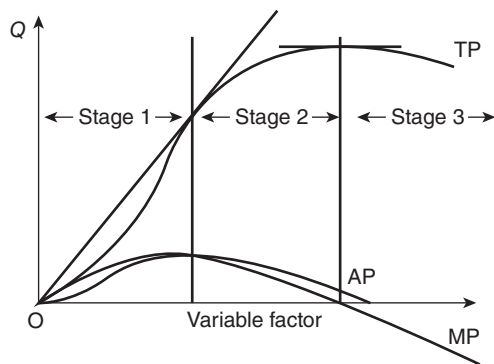


Figure 1.3 Total, average, and marginal product curves.

equals marginal product (MP). Figure 1.3 depicts the average product and marginal product curves corresponding to the total product curve and shows the intersection of the AP and MP curves below the intersection on the TP curve of the line from the origin that has the greatest slope. Figure 1.3 also marks out three stages of production on the basis of the relationship between average and marginal product. In Stage 1, the average product of the “variable factor” is increasing. Symmetrically, the marginal product of the “fixed” factor is negative. The boundary between Stages 1 and 2 is the maximum point of average product. In Stage 3, marginal product of the variable factor is negative. The boundary between Stages 2 and 3 is the point of maximum total product, indicated by the horizontal line tangent to TP. Producing at any ratio of the variable factor to the fixed factor contained in Stage 1, the producer could get a larger average product by adding more of the variable factor, and he or she would be irrational not to add more of the variable factor. Consequently, production in Stage 1 is irrational. In Stage 3, the producer has added so much of the variable factor that the units are literally tripping over one another; they actually lower total product, which is the meaning of a negative marginal product. Production in that stage is also irrational. Stage 2 contains the only ratios of factors (inputs) that it is rational to employ. One of the thoughts to take away from this exposition is that producers will always produce in a range (of input ratios) of decreasing marginal product, for all inputs. Explanations of people’s actions as being efforts to get away from, or avoid, decreasing marginal productivity are incorrect.

In Book XI of *De Re Rustica*, ll. 17–18, Columella notes that a specific area of land, an *iugerum*, can be trenched for a vineyard to a depth of 3 feet by 80 laborers working for one day, to 2½ feet by 50 laborers, or to 2 feet by 40 laborers. Notice the constant marginal returns, in terms of depth dug, between the application of 40 and 50 laborers and the decreasing returns when he increases the number of laborers to 80: 80 laborers can dig less than twice as deep as can 40 laborers (Forster and Heffner 1955, 79). This, of course, is not an empirical observation but, possibly even more important, it is a recognition, or expectation, of decreasing marginal returns to

increasing applications of labor to a fixed quantity of land.

1.3 Substitution

The next technological relationship specified by the production function that we will discuss is the array of ways that different combinations of the inputs (two in this case) can produce a given quantity of the output. You also can think of this topic as how the inputs relate to one another. In Figure 1.4, the quantity of labor (N) is measured on the abscissa (the horizontal axis) and the quantity of land is measured on the ordinate. The curved line labeled Q_0 represents a constant quantity of output, say 100 bales; it can be produced with any of the combinations of land and labor represented by coordinates lying on it. Thus, the labor-land combinations represented by A (N_0, L_0) and B (N_1, L_1) will both yield 100 bales of cotton (Q_0). The curve Q_0 is called an *isoquant*, because each point on it represents the same quantity of output. Isoquant Q_1 represents a larger quantity of cotton, say 200 bales. Combinations of labor and land represented by points C (N_3, L_3) and D (N_1, L_4) will both produce 200 bales of cotton. Notice that, as these isoquants are drawn, it is not necessary to use larger quantities of both inputs to produce a larger output; in fact, we can produce 200 bales at point D using no more labor than we used at point B to produce 100 bales (N_1) if we are willing to increase our use of land to L_4 from L_1 . This concept ("there's

more than one way to skin a cat," begging my own cats' pardon for the expression) is known as "substitution." Specifically, the production function represented by the family of curves Q in Figure 1.4 indicates that there is substitutability between land and labor in the production of cotton. Empirically, most production technologies embody substitutability between (among) inputs.

The alternative – nonsubstitutability – can be represented graphically as the L-shaped curves in Figure 1.5. We can combine N_0 units of labor and L_0 units of land to produce Q_0 units of output. If we add some labor, say to N_1 , but keep land unchanged at L_0 , we still get Q_0 units of output, so we just wasted labor in the amount $N_1 - N_0$. Only land-labor combinations along the line labeled R will be efficient; above R , we're using land that contributes nothing to output, below it we're using labor that contributes nothing. Such a production technology commonly is called a "fixed-coefficients" technology. Why even consider a production function with such a characteristic? Several reasons. First, it is one logical end of the continuum of degrees of substitutability between inputs. Second, for very short periods of analysis, in which it is difficult to substitute among inputs, many technologies with flexibility over longer periods can be studied as if they were fixed-coefficient technologies. The technique known as input-output analysis generally specifies fixed-coefficients technologies.

Let's return to the isoquant diagram and the issue of substitutability among inputs. Figure 1.6

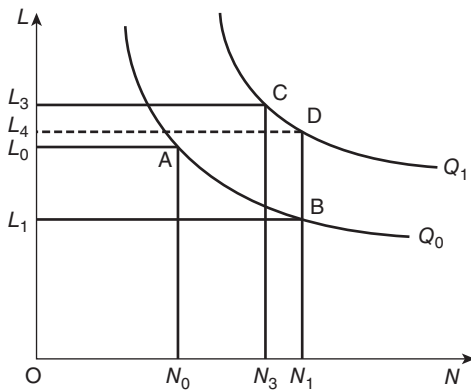


Figure 1.4 An isoquant with substitution between inputs in the production technology.

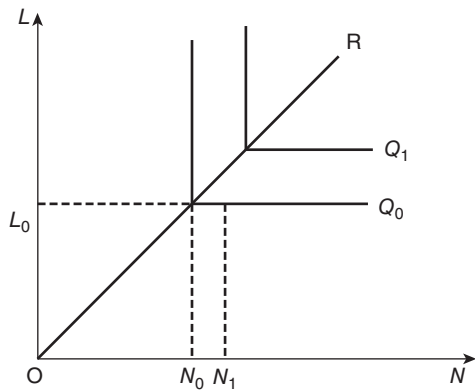


Figure 1.5 An isoquant with no substitution between inputs in the production technology.

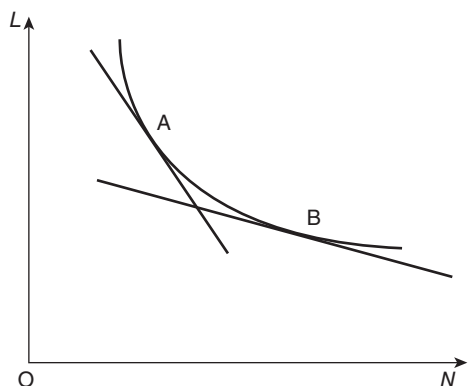


Figure 1.6 Marginal rate of technical substitution (MRTS).

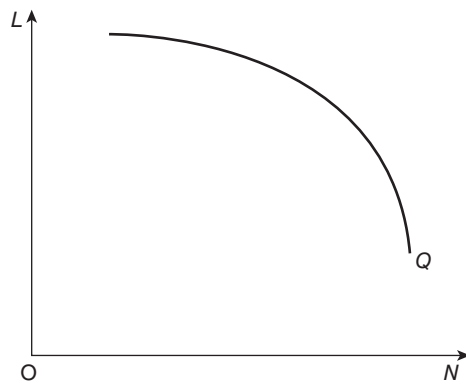


Figure 1.7 An isoquant with increasing MRTS – impossible.

reproduces isoquant Q_0 with points A and B from Figure 1.4. The two lines drawn tangent to points A and B have marginal interpretations analogous to the tangent to the total product curve (TP) in Figure 1.2. The slope of the line tangent to the isoquant at point A represents the number of units of land (L) that have to be substituted for a single unit of labor (N) at that point (the change in L divided by the change in N). The slope is steep relative to the slope of the line tangent through point B. Point A represents a labor-land input combination that uses relatively few units of labor. At such a point, substituting even more land for another unit of labor is relatively difficult. At a labor-land combination like point B, where the ratio of labor to land is high, substituting a unit of land for labor is not nearly so difficult. The slope of the isoquant (actually, the negative of the slope) at any point is called the *marginal rate of technical substitution* (which itself is, in fact, the ratio of the marginal products of the two inputs at that ratio of inputs; we will discuss the concept of the marginal product shortly).

The reader may have wondered why the curvature of the isoquant that allows substitution between inputs is shaped the way it is. Specifically, why is it convex, as Figure 1.4 shows, rather than concave, as in Figure 1.7? We have already presented the information to answer this question, but it may be useful to reassemble it here. The convex isoquant of Figure 1.4 indicated

diminishing marginal rates of technical substitution as we moved toward either axis. That is, as more labor is substituted for land (to the right end of the abscissa, or N -axis), it takes progressively more labor to replace a unit of land and still produce a constant output. Viewing this corner of Figure 1.4 alternatively (moving from right to left instead of from left to right), when we are already using a lot of labor, the amount of land required to replace a unit of labor and keep output constant isn't very large. If we had a concave isoquant such as Figure 1.7 shows, our technology would be characterized by increasing marginal rates of technical substitution: as we replaced more land with labor, we could substitute away units of land more and more easily. As we will see below when we introduce the role of input prices in determining input ratios in production, a concave isoquant would encourage the use of higher proportions of the relatively more expensive input.

Figure 1.8 shows an isoquant that possesses infinite substitutability between land and labor. At any location along the isoquant, a unit of land can substitute for x units of labor (where the value of x is determined by the slope of the isoquant). Perfect substitutability does not play a large role in economic analysis, probably because it is not important empirically. We present it simply to show the limiting case of substitutability in production.

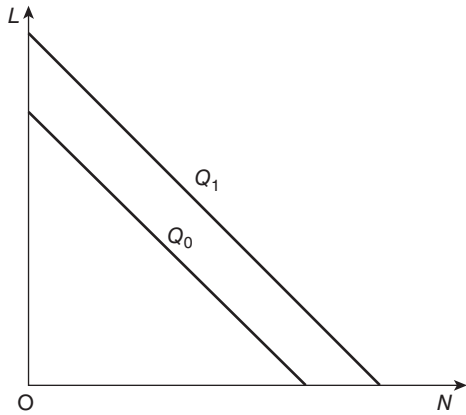


Figure 1.8 An isoquant with perfect substitutability between inputs – unlikely.

Much of the practical agricultural advice contained in the Roman texts such as Columella's *Res Rustica* and *De Arboribus* and portions of Pliny the Elder's *Natural History* is written as if the combinations of resources used in various crops and husbanding were required in very specific proportions, very much as would be implied by fixed-coefficients production functions. Nonetheless, even in these texts we can find discussions of alternative ways of doing things. Pliny, in Book XVIII of the *Natural History*, l. 35, notes that, at least in older times, it was considered better to sow less land and plough it better – clearly a substitution of labor for land (Rackham 1950, 213).

1.4 Measuring Substitution

Recall from Figure 1.2 that we can calculate the marginal products of both inputs – and consequently the ratio of their marginal products – from knowledge of the ratio of the quantities of the two factors (with, of course, knowledge of the “functional form” of the production function, which we will discuss below). A summary measure of the degree of substitutability between inputs in producing a constant quantity of output, called the *elasticity of substitution* (between

inputs), is the percentage change in the ratio of inputs divided by the percentage change in the ratio of marginal products. It is always measured positively; frequently the lower case Greek letter σ (or σ_{ij} – read as “sigma-sub ij” – for the elasticity of substitution between inputs i and j when there are more than two inputs in the production function) is used to denote it. In a more mathematical treatment than we will use here, there are a number of ways of deriving formulae for the elasticity of substitution, some using strictly characteristics of the production function, others using input prices; none is “wrong,” but different measures illuminate different aspects of substitution and different circumstances. Another, fairly intuitively appealing formula defines the elasticity of substitution between two inputs as the negative of percentage change in the ratio of the quantities used divided by the percentage change in the ratio of their costs. The elasticity of substitution – indeed any elasticity – is a pure, dimensionless number. That is, it does not have the dimensions of output/input or cost/quantity, or whatever; it will have the dimensions of input/input or cost/cost, such that the measured units cancel. (If, in modeling some problem yourself, you find occasion to construct an elasticity and you find that it has the dimensions of, say, distance over time, or some such, you’ve made an error.)

When a production function has only two inputs, those inputs are always substitutes for each other. In the cases of three or more inputs it is possible for some pairs of inputs to be complements. In the case of substitutes, when the relative price of one input goes up – call it input A – the ratio of input A to substitute input B would fall as the producer substitutes B for A . If inputs A and C are complements to each other, when the use of one of those inputs falls because of a rise in its relative price, the use of the complement also will fall; whether the ratio of the two complementary inputs falls, rises, or remains constant is an empirical matter. Nevertheless, the ratios of both those inputs to input B , for which they must be substitutes, will fall when the price of one of them rises relative to the price of B . The issue of substitutability or complementarity is important

in the subject of the demand for inputs, which we will discuss below.

1.5 Specific “Functional Forms” for Production Functions

Since we have brought up the concept of the “functional form” of a production function, let’s discuss it somewhat further. We introduced the concept of the production function with general, functional notation $f(\blacksquare, \blacksquare)$, where “ f ” (it could have been any letter, Roman, Greek, or otherwise) deliberately avoids spelling out exactly what the equation looks like. Recalling junior high school algebra, some function $y = f(x)$ could represent a specific equation like $y = a + 2x$, where x is the “independent” variable and y is the “dependent” variable. (On a Cartesian graph, such as we’ve used here to describe the behavior of production functions, y is on the ordinate and x is on the abscissa.) Several specific functional forms have been extremely popular for production functions, because of both their theoretical properties and their ability to find empirical correspondence in data on production.

The simplest functional form that allows substitutability between (among) inputs is the Cobb–Douglas function: $Q = AN^\alpha L^\beta$, in which A is simply a constant term, which turns out to be handy to represent such events as technical change. First, note that if the value of either input (N or L in our cotton case) is zero, the value of Q will be zero. The exponential parameters α and β , called “output elasticities,” are positive and generally add up to a value close to 1.0. We’ve run into the term “elasticity” already, in reference to substitutability. Elasticities are widely used in economics to describe the percentage change in one quantity (the one in the numerator of the ratio) caused by a 1% change in another quantity; the elasticity is the percentage change in the “dependent” variable divided by the percentage change in the “independent” variable. An output elasticity is the percentage change in output attributable to a 1% change in the corresponding input. The sum of the output elasticities in the Cobb–Douglas function has an important physical interpretation: it is the degree of returns to

scale in production. A sum of output elasticities exactly equal to 1.0 implies constant returns to scale (sometimes abbreviated CRS): a 1% increase in all inputs will yield exactly a 1% increase in output. A sum of output elasticities greater than 1.0 implies increasing returns to scale, and a sum less than 1.0 gives decreasing returns to scale. An example of increasing returns to scale would be if a 1% increase in all inputs yielded a 1.05% increase in output. For decreasing returns to scale, a 1% increase in all inputs would yield, say, a 0.95% increase in output. A restrictive feature of the Cobb–Douglas function is that its elasticity of substitution between each pair of inputs is exactly 1.0, and the elasticity of substitution has exactly that value at all points on the isoquant. (As such, the Cobb–Douglas function is one of a class of production functions called “constant elasticity of substitution” functions. This is in contrast to production functions that allow the elasticity of substitution to vary at different points along an isoquant, an apparently “nice” characteristic when one wants to study the effects of substitutability quite closely but one that adds enormous mathematical complexity to any analysis.) Consider the magnitudes of the output elasticities α and β . Under CRS, reasonable values of these two parameters would be $\alpha = 0.8$ and $\beta = 0.2$. By “reasonable,” we mean that considerable empirical investigation of agricultural production with the Cobb–Douglas production function has yielded statistically estimated values of closely equivalent parameters around this pair of values. Now, what does it mean to say that the output elasticity of labor is 0.8? A 1% increase in the use of labor, holding constant the amount of land used, will increase output by 0.8%. Doubling your labor alone *won’t* double your output: such a proposition ignores the fact that labor isn’t the only thing that contributes to the production. It would increase it by 80%. Correspondingly, increasing your land by 1% would increase output by 0.2%; doubling your land input would get an additional 20% of your output.

Another especially popular functional form for production functions is the so-called constant elasticity of substitution, or CES, function: $Q = A[\delta N^{-\rho} + (1-\delta)L^{-\rho}]^{-1/\rho}$, where the elasticity of substitution between inputs N and L is

$\sigma = 1/1 + \rho$, and the value of ρ is between positive infinity and -1.0 . The A term is comparable to the A term in the Cobb–Douglas function. The parameter ν indicates the returns to scale ($\nu = 1.0$ for CRS). The δ coefficients represent the intensity of use of the inputs, but are not exactly comparable to the output elasticities of the Cobb–Douglas; in fact the output elasticities for the CES function are quite complicated formulae rather than single parameters. The CES is a much more difficult functional form to use for analytical (as contrasted with empirical) study. Nevertheless, this functional form allows the elasticity of substitution between each pair of inputs (all elasticities are constrained to be the same value) to be greater or less than unity, which can have significant implications for the demands for inputs as their relative costs change. (We have not discussed demands for inputs yet – or demands for products for that matter; the concept, applied to inputs, describes how much of the input a producer will want to use, according to its productivity and cost. The issue is of critical importance in determining the distribution of income in an economy among the owners of various factors of production.) When the elasticity of substitution in the CES function is unity ($\sigma = 1.0$ when $\rho = 0$), the form collapses to the Cobb–Douglas form; when $\sigma = 0$ (as $\rho \rightarrow \infty$; in other words, “goes to infinity”), it collapses to the fixed-coefficients production function.

Considering the limitations of these two production functions, we have to say a few words explaining why they maintain their popularity. Contemporary empirical (econometric) study of production favors more sophisticated functions, such as the transcendental logarithmic (“translog”), which allows any degree of substitutability (or complementarity) between any pair of inputs and allows substitutability to vary along isoquants. This functional form has a large number of parameters, which requires a correspondingly large data base for statistical estimation. In circumstances where data are less readily available, the CES and even the Cobb–Douglas are still used. In analytical uses (just writing equations and diagrams with pencil and paper), both the Cobb–Douglas and the CES can demonstrate many interesting theoretical

issues while offering considerable mathematical tractability (particularly the Cobb–Douglas). The translog function would be quite difficult to manipulate for heuristic purposes, and would offer little in the way of additional insights to compensate for the greater trouble. The engineering production functions we introduced in section 1.1 generally are far more intricate than any of these functional forms designed for analytical or empirical research.⁴

1.6 Attributing Products to Inputs: Distributing Income from Production

After this brief excursion into functional forms, let's return to the issue of marginal products of inputs. We've seen that the marginal (physical) product (MPP) of an input is the contribution that an increment of the input makes to total output. Under conditions of constant returns to scale, total output can be decomposed into a sum of MPPs: in our case of producing cotton with labor and land, $Q = \text{MPP}_N N + \text{MPP}_L L$. Now, think of the cost of producing Q : we have to pay for labor and land. Let's put the cotton in terms of its value by multiplying the entire equation by the price of cotton, p : $pQ = p\text{MPP}_N N + p\text{MPP}_L L$. Now, thinking in terms of “wages” and “rents” for labor and land (terms to which we will return shortly), we can express the revenue from the cotton we produced as $pQ = wN + rL$. The wage rate (or the “price” paid for labor, by any other name) is equal to the marginal physical product of labor (which is actually in cotton) times the price of cotton; and similarly for the rental rate (or the “price” paid for using land this season). If we were working in a barter economy (that is, one in which money doesn't exist and people purchase one good directly with another), the payments to labor and land (or to the people who own those factors of production) are made directly in the output, cotton. (What happens to this simple equation when there are either increasing or decreasing returns to scale? With decreasing returns to scale, payment according to marginal productivity will more than exhaust the output – that is, there won't be enough to go

around; with increasing returns to scale, there'll be product left over after paying all the factors their marginal products. This doesn't cause as severe a problem for marginal productivity theory of factor pricing – and the income distribution theory based on that – as it might seem, but we'll have to come back to why.)

We can obtain more information out of this cost relationship. We can divide our cotton revenue-cost equation by the value of the cotton output to get an equation in terms of cost shares: $1 = wN/pQ + rL/pQ$, where wN/pQ is the proportion of the cost of cotton production that can be attributed to labor and rL/pQ is the proportion attributable to land. These are commonly called “cost shares” or “factor shares.” However, it can be shown mathematically that these cost shares are equivalent to the output elasticities of their respective inputs: the percentage change in output divided by the percentage change in input, or the ratio of the marginal product to average product of each input. Recall that w , the wage rate, is the marginal physical product of labor, times the price of the output, p ; since we have w/p , the p s cancel and we're left with just the marginal physical product of labor. This is multiplied by N/Q , which is one over the average product of labor; so the entire “share” expression is the marginal product of labor divided by the average product, which is the definition of the output elasticity of labor in the production function.

Having introduced the concept of the factor share, this is a good place to note that the elasticity of substitution gains particular interest for its role in determining the distribution of income among the owners of factors of production. Suppose for the moment that we have two principal factors in our economy (or at least in our model of our economy) – labor and land – and that our economy produces a single good – food. An abstraction, admittedly. If the elasticity of substitution between land and labor in the food production function is unity (1.0), a change in the relative price of land and labor, caused possibly by technological change, population growth, expansion of arable, or some other major event, will leave the factor shares unchanged. However, if $\sigma > 1$, the share of the factor whose relative price has fallen will increase at the expense of the other factor. For example, with $\sigma = 1.5$, say, if the

relative price of land falls, land will be substituted for labor to an extent that the relative share of total income going to land will increase; since there are only two factors, that of labor will fall. If $\sigma < 1$, the relative income share of the factor whose relative price has increased will rise at the expense of the other factor.

1.7 Efficiency and the Choice of How to Produce

Let's return to our isoquant version of the production function. Why should we pick one point on it for our input combination rather than any other? In Figure 1.6, the slope of the isoquant at any point represented the rate at which we could substitute land for labor (or labor for land) and still produce the same amount of output. That described our technological capabilities. The negatives of sloped lines in that diagram also represent the cost of land in terms of labor – either minus the rental rate on land divided by the wage rate of labor if we want to use a monetary numeraire, or the number of units of land we could rent if we were to trade a unit of labor for it in the case in which there is no money to use for a numeraire. Either way – with money or without – the (negative of the) slope of a line “in $L-N$ space” represents the availability of land and labor to our producer. The isoquant represents the technical ability to substitute land for labor and still produce the same output, and a “price” or “cost” line represents our producer's ability to secure the services of those two inputs. At a point of tangency between such a price line and an isoquant, the producer can substitute between labor and land in production at the same rate at which he or she can “hire” or “rent” them. In general, higher costs of land relative to labor will prompt producers to use higher ratios of labor to land; similarly for ratios of any two inputs in proportion to their relative costs.

This description of the conditions of efficiency in production may sound fine as theory, but it is legitimate to ask how real people might discover such efficient allocations of their resources for themselves. First, agents directing production operations for themselves or for others can be expected to have a good, first-hand idea of what

their input costs are. Even if they do not hire inputs on an open market in an easily measured numeraire such as money, they can be expected to have a good, working idea of what they would have to pay in kind or cash for additional units of each of their inputs. Next, how do they find out about the rates of technical substitution in their production technologies? Two ways: experience and the pressures of competition. Experience is self-explanatory by and large. Competition can come from the interactions of a large number of other individuals interested in bidding away resources for other activities or in supplying the same products as our agent under consideration. Alternatively, staying a step or so ahead of the grim reaper (competition with nature) can have a similar effect in, as Dr. Johnson expressed it, concentrating the mind wonderfully. Does this mean that all societies at all times are perfectly efficient? The answer is, naturally and obviously, “No,” but neither can they be expected to leave a lot of so-called “low-hanging fruit” around to rot. Efficiency in any real conditions depends on the users’ understanding of their technology and, to some extent, on their understanding of how their own societies operate and respond to opportunities and incentives.

It is important for students of economies, ancient and modern, to distinguish between efficiency and productivity. Ancient agriculture used low-productivity technologies, but chances are excellent that ancient farmers used those low-productivity technologies highly efficiently. The ancient land transportation industry similarly is invariably characterized as inefficient, a quite unlikely state of affairs. Efficiency is a matter of how close the marginal rate of technical substitution (along an isoquant) is to the marginal rate of substitution of inputs as represented by a relative price line in our diagrams or, more generally, by producers’ ability to acquire an extra unit of one input in exchange for some quantity of another input. Productivity is represented by how far from the origin of our diagrams an isoquant representing a particular quantity of output is located: a unit isoquant (representing the quantity of inputs required to produce one unit of output) closer to the origin uses fewer inputs than one farther away, hence representing greater productivity. Efficiency refers to the behavioral choice of *where* on that isoquant to

produce – that is, given a relative price of inputs and the input substitutability within a technology, how close to the maximum possible output the producer gets from his resources. The difference in contemporary scholars’ attitudes toward the people of antiquity, depending on whether we view them as having been inefficient – with all the other pejorative characteristics associated with that unfortunate state of being – or efficient but burdened with unproductive technologies, could have broad consequences for our own studies.

Economic efficiency is not a product of the modern, industrial world, but is simply getting the most out of one’s resources that one can, subject to the institutional constraints one faces. In Chapter 6, we’ll discuss the role of constraints in modifying an absolute efficiency concept to various forms of conditional efficiency. For a consumption-oriented example, the absence or poor development of information markets to support the Roman housing market, as noted by Frier (1977),⁵ probably did retard the rapid matching of people wanting to occupy housing with those having units available, but information is a tricky good to produce, economically speaking, as we will learn in Chapter 7. Given the limited information available on housing, there is little reason to suspect that people knowingly made less of their resources in housing than they believed they could. In pursuing the issue of inefficiency in the Roman housing market further, the tendency to execute long-term contracts and the institutionalized payment after occupancy rather than before or during both could be ascribed to the limited production of information. Introducing concepts from four subsequent chapters in the quasi-empirical discussion of efficiency is not a deliberate tease, but rather a demonstration of the intricacy of the empirical application of the efficiency concept. When ancient institutions supporting some activity do not demonstrate the same capacities of flexibility and overall productivity that typically accompany corresponding activities in the post-World War II period in the Western, industrialized nations, it is simplistic, as well as just plain wrong, to adopt the fallback position that those people did not act economically or that their activities were simply governed by social restraint. Better to investigate the economic reasons for the ancient

constraints, as Stambaugh has done regarding the public services that were and weren't offered in Roman cities.⁶

1.8 Predictions of Production

Theory 1: Input Price Changes

Let's exercise the theory a bit, using this last set of relationships about picking the optimal input ratios according to the prevailing price or cost ratios. Figure 1.9 has a lot of lines in it, but we can walk through them and take away the information they convey. The production technology is characterized by the family of isoquants Q_i , of which we have drawn just three. The amount of output associated with the isoquants increases as we move outward from Q_0 to Q_3 . We begin with the situation in which the relative price of land and labor is characterized by line AA' , which is tangent to isoquant Q_0 at point 1. Our producer (this "producer" might be an individual, a firm, a family farm, a temple, or an entire region or country) finds that it can produce the most output with its technology by using L_1 amount of land and N_1 labor. The line from the origin, R_A , is called an expansion path; it describes the combinations of land and labor that this technology would employ if it were to expand at the constant set of relative prices described by line AA (refer to A' as "A prime"). Let's consider a change in this situation: the relative price of labor drops

from AA' to AB . But before we proceed, how do we know that such a counterclockwise pivot of the price line around its intersection with the ordinate (the land axis) represents a cheapening in the relative cost of labor? Here's one way. Suppose that the actual intercepts of price line AA with both axes represent the real resources available to the producer: if the producer decided to put all available resources into the acquisition of land and none into hiring labor, OA is the quantity of land that could be acquired (rented) at the relative prices described by AA' . Alternatively, if she were to devote all her resources to hiring labor at the same relative price ratio, she could hire the services of OA' labor. (There's no good reason why any producer would want to put all resources into just one input; this is just a method of demonstrating a point.) Now, the relative price changes to the line AB . With the same resources, the producer could still rent OA units of land but could hire OB units of labor, which is considerably more than she could hire under the relative prices of AA' . Consequently, labor is cheaper relative to land under AB than under AA' .

Now, the relative cost of labor has fallen, and the production technology has remained unchanged. The highest isoquant our producer can reach with the resources characterized by the intercepts of relative price line AB is Q_2 . The movement from the input combination (L_1, N_1) to input combination (L_3, N_3) includes a substantial decrease in the ratio of land to labor represented by the shift from expansion path R_A to expansion path R_B . This move includes both a substitution effect and a scale-of-production effect. If we were to change the relative price from AA' to AB but restrict the producer to the same level of production, the input combination would still move toward more labor and less land; the same relative price of AB is reproduced in $A''B'$ (refer to A'' as "A double-prime"), which is tangent to Q_0 at point 2. Here the producer uses less land than before ($L_2 < L_1$) and more labor ($N_2 > N_1$), but still produces the same amount of output. Since we're letting the change in the relative price reflect a real change in the resources available to the producer, she can expand her scale of production to the point where some isoquant will be just tangent to the new relative price line AB . The

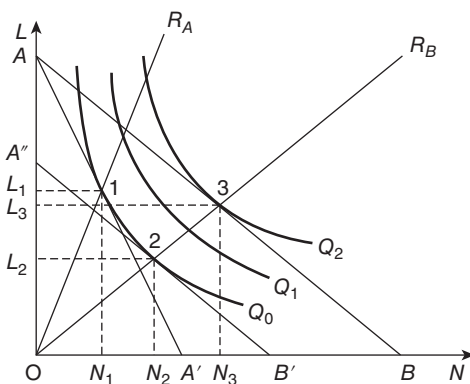


Figure 1.9 Production responses to input price changes.

producer *could* produce only the amount of output described by isoquant Q_1 , but she is able to reach as great a scale of production as that associated with Q_2 . The movement from point 2 on Q_0 to point 3 on Q_2 represents the scale effect; land used rises from L_2 to L_3 , and labor hired rises from N_2 to N_3 . This change in relative price could represent an actual change over time (or even instantaneously) in a single location or a comparison of production choices involving the same technology but different locations with different resource availabilities.

Let's consider a couple of applications of this concept. Agricultural technology in Egypt and Lower Mesopotamia during the middle of the second millennium had much in common, to avoid saying outright that it was identical. Similar arrays of crops were grown with pretty much the same array of tools and animals, and with comparable biological understanding on the parts of the two societies' farmers. Water supply in Egypt was primarily by inundation, with various lifting equipment, while the Mesopotamians supplemented with a more extensive system of canal irrigation, supplemented with comparable lifting equipment. The different water-supply systems may have altered the price of water relative to other inputs, such as seed, labor, animal traction, and hand-held equipment between the two regions. Different population densities would have altered the relative availabilities (and hence costs) of labor and land. We can expect that these differences in relative prices would have had some impacts on the ratios of a number of these inputs, making Egyptian and Mesopotamian agriculture look more different than they actually were at a fundamental, technological level.

Correspondingly, dry-land agriculture in Upper Mesopotamia, with its different relative cost of water (relative to the other inputs such as labor, land, and equipment) than existed in Lower Mesopotamia, would have conferred a considerably different appearance to agricultural practices in the two regions. The seed/land ratios would have responded to the land/water cost ratio, and if more plentiful availability of water enhanced the value of land in Lower Mesopotamia, we would expect to have seen lower ratios of labor to land in Upper Mesopotamia.

1.9 Predictions of Production

Theory 2: Technological Changes

Consider another possible change or difference. Figure 1.10 can represent either a technological change facing a given producer or a difference in technologies faced by producers at different locations. Using the technology associated with isoquant Q_0 , a producer facing relative prices represented by AA would choose input combinations along expansion path R_0 . Facing the same relative prices but using a different technology, represented by isoquant Q_1 – in which the substitution of labor for land is more difficult at each land-labor combination – a producer would use a higher ratio of labor to land, represented by input choices along expansion path R_1 .

Consider an example of technological change later in antiquity – the Roman use of pozzolana for a hard, strong cement. In addition to the possibility of producing entirely new products (structures) such as true arches, the increased material strength would have permitted the substitution of land used in structures to actual construction material: buildings would have been able to cover larger floor spaces because the relative price of material strength to land's price had fallen. Additionally, the ratio of usable space per unit of land would have increased as the relative price of its provision fell.

Not all changes in the way things are done are technological changes. Some observations

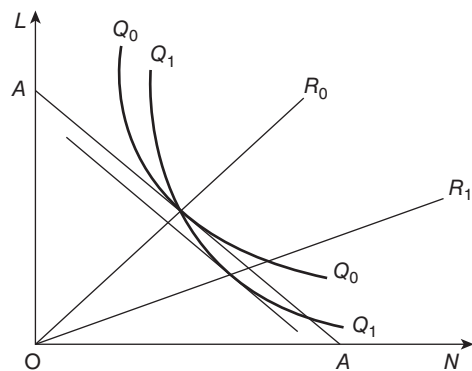


Figure 1.10 Production responses to a change in production technology.

in Laurence's (1999, Chapter 5) recent study of Roman roads can be used to illustrate this point. Laurence notes that the apparent growth in the use of basalt (*silex* or *selce*) in road paving from the third century B.C.E. into the first century C.E. may have been limited primarily (with the glaring exception of the Via Domitiana) to short stretches of "showcase" roads and some city streets. He does not attempt to attribute any part of the increase in its use to technological improvements in quarrying this stone, and indeed it is difficult to see the increase in use of *selce* as representing a technological change. As Laurence notes, the weathering properties of this stone may have made it more conducive to foot and pack animal traffic than to wheeled vehicles (Laurence 1999, 72). Most roads in Italy continued for some time to be surfaced with compacted gravel (*glarea*), but Laurence notes that these roads appear to have been gradually upgraded in quality: "the nature of the road surface and its associated structures were upgraded and altered to reflect changes in the available technology, with a marked improvement in terms of the speed of travel or the weight of goods... that might have been transported" (Laurence 1999, 73). It is not clear that the gradual upgradings cited here were indeed permitted by improvements in construction or materials technology, as opposed to having entailed simply more extensive applications of the same technology in response to demands for more roads to accommodate the growth of traffic, in terms of both flow volumes and weight of goods carried.⁷ It is possible that technological changes in vehicle technology and tackle for the animal prime movers could have contributed to increased demands for road durability, width, and speed qualities, but no evidence is suggested that changes in actual road construction technology occurred in these particular upgrades. This is not to imply that technological improvements in Roman road construction did not occur, such as modifications of substructural support, use of sand and pebbles for grading under paving stones, and changes in preferred materials because of their properties of cementation, durability, flexibility, and so forth – changes that would have altered isoquants for roads. The advances in bridge

construction that permitted wider valleys to be spanned (Laurence 1999, 73–75) probably do represent what would be represented by changes in an isoquant (quantities and proportions of, and substitutabilities between, labor, equipment, and materials required to build a bridge of specified dimensions). This discussion is not a semantic or nomenclatural cavil, but an emphasis of the distinction between a change in the isoquant involved in constructing particular types of infrastructure and growth in the quantity and quality of infrastructure, because of an increased demand for it, using a constant isoquant. The importance of the distinction should be obvious: first, no one would want to confuse growth without technological change for technological change; and second, the implications for who benefitted from the observed events are different as the incomes of the owners of different factors would have fared differently under the alternative circumstances.

1.10 Stocks and Flows

In our discussion of factors, we have referred to "renting" land, "hiring" labor, or the ambiguous term "acquiring the services" of either. It is intuitive to visualize the productive input called "labor" as individual people and to think of their use in a productive activity as occupying their entire persons. Correspondingly with land: why not just "buy" it and "use" it as much as you want? We have deliberately avoided including physical capital, or items of equipment, among our factors of production, for reasons to be discussed below, but the same issue arises very clearly with equipment. If we use a hammer among our inputs, what is the relation between "owning" the hammer and "using" it?

For each type of input, it is useful to distinguish between the stock of the input and the flow of services derived from a unit of it in each time period. The form in which a stock of labor appears is the individual person; even if the person happened to be owned by the agent organizing the production (a case of slavery), only the *services* of the person are used in production during any period. Moving to a less potentially controversial

case, consider the hammer we just introduced. Treated reasonably carefully, a hammer will last several periods; we can use it this period, next period, the period after, and so on. In each period, we use the services of the hammer. The services are derived from the “stock” of hammer, and it might be possible to eventually “use up” the stock of hammer – i.e., either wear it out gradually or break it all of a sudden, just through regularly conscientious use. Consider land. Land sometimes is (was) thought of as an “original and indestructible” factor. However, land can be “overused” and its fertility exhausted. Farmers routinely conduct maintenance of one kind or another on their land to keep it from washing away through erosion, burning out through salt accumulation, or otherwise becoming less productive.

Production theory works with flows of services of factors during a given period of time. These flows are derived from a stock that embodies the factor. The acquisition cost of a flow of services from a factor for one time period generally is substantially lower than the acquisition cost of the stock of the factor. It will be helpful to maintain a conscious distinction between stocks and flows in many contexts; ignoring or confusing the two concepts can lead to serious analytical errors.

1.11 The Distribution of Income

We should say a few words about the distribution of income at this point. First, what do we mean by the distribution of income? There are two principal interpretations of this expression, the functional and the personal distributions. The latter is possibly more intuitive: it refers to how total income in the economy is distributed among individuals or families. Frequently it is measured by the Gini coefficient, whose numerical values can be interpreted as degrees of skewness. For instance, in a number of Latin American countries in the 1950s and 1960s (and later as well), the 1 to 2% of families with the highest incomes in the country received around 20 to 30% of total national income, and the top 10% about 50%, while the bottom 60% received around 20% (Jain 1975, 24 Table 13, 89 Table 57; Fishlow 1976,

61 Table 1, 72 Table 5; Webb 1976, 12 Table 1, 13 Table 2; Weisskoff 1976, 35 Table 1, 38–39 Table 2). They also held a similar proportion of national wealth (a stock-flow distinction). The 20% of families receiving the lowest incomes pulled in about 5% of national income. Together with the people somewhere in the middle, these numbers represent the personal distribution of income in these countries. The personal distribution of income has considerable political importance, as it is easy to imagine. These countries may be a reasonable benchmark against which to gauge personal income distribution in antiquity.

The functional distribution of income describes the proportions of total income in the economy going to particular factors of production (land, labor, capital), or more specifically, to the owners of those factors. A functional distribution of income underlies each personal distribution of income: the factors still produce the income, regardless of who owns them. Following the neoclassical model of the functional distribution of income, the factor shares (proportion of input costs accounted for by each factor of production) from individual production functions can be aggregated across the economy to reach aggregate factor shares, which amounts to the functional income distribution. One of the very handy features of the way we have studied production is that, under constant returns to scale, the output elasticity of each factor (of production) equals its share of income from a production process. Proceeding to the level of the aggregate economy, it is not difficult to find that, with a land output elasticity of 0.15, a capital output elasticity of 0.05, and a labor output elasticity of 0.8, labor claims 80% of total income in the economy, the owners of land 15%, and the owners of capital 5%. These output elasticities are characteristic of the output elasticities of “traditional” agriculture (i.e., the agricultural sector that uses animal and human power rather than fossil fuels and natural rather than chemical fertilizers) over the past several decades. Of course, if 5% of families end up with 40% of income, we need to look into how some of the labor income of the bottom 95% of families is being captured by the 5%.

The neoclassical theory of income distribution has been criticized from several directions, primarily for its use of the construct of aggregate capital rather than a plethora of individual items of equipment and because “market imperfections” (a term we have not discussed yet, generally used to refer to departures of industry structure from that of perfect competition; see Chapter 4) cause the incomes to factors to differ from their values of marginal product (which equality is what lets us associate the output elasticities with factor shares). The most important of these disagreements about neoclassical income distribution theory is known in the economics literature as the Cambridge Controversy, or sometimes the Cambridge–Cambridge Controversy; the leading critics of neoclassical income distribution theory have come from Cambridge University, in England, while its most cogent defenders were at the Massachusetts Institute of Technology, in Cambridge, Massachusetts in the United States. We will not devote much space to that discussion; the most telling criticisms of neoclassical distribution theory appear to have their strongest force in situations that are not particularly important empirically, and abandoning the simplicity and power of the neoclassical theory would leave us effectively without an alternative theory of functional income distribution. So, as is the case with many theories in science, while it may not be perfect, it will get us by until a superior theory emerges, which to date has not occurred.

It is easy to see that neoclassical distribution theory relies on both supply and demand influences to arrive at a distribution of income. Output elasticities in production functions are clearly and purely technological parameters. However, it is the value of marginal product that determines factors’ shares; value of marginal product of any factor is the price (value however expressed) of the product times the marginal physical product of the factor – the physical amount of what it produces. Prices reflect individual and group valuations – that is, the foundations of product demand, which we discuss in Chapter 3.

Using neoclassical income distribution theory gives us a baseline, functional distribution of income that “should have” appeared in many of the ancient Mediterranean and Aegean

economies, considering their production technologies. To the extent that we can get occasional glimpses from either textual or artifactual records that the personal distribution of income (which may be easier to observe for those places during the long periods of antiquity) that differ substantially from the 15-5-20 distribution we noted in the previous paragraph, there is a need for explanation. Neoclassical income distribution theory gives us an implicit baseline that needs explanation. Some combination of taxation, tribute, slavery, and imperfect markets (monopolization of some productive activities is one such imperfection) are obvious candidates. More subtle possibilities derive from other market imperfections, including the possible absence of markets for such items as insurance of various sorts (see Chapter 3, on consumption, for a discussion of risk and insurance, although there is no treatment there of the absence of insurance provision). Other theories of the personal distribution of income would assist in this explanation, but since some of the most incisive of those theories rely on concepts we have not introduced yet, we defer further discussion of them to a later chapter.

Before leaving the subject here, however, we offer a brief preview of what to expect in the way of the analytical treatment of income distribution. So far, we have treated production in a partial-equilibrium approach. The alternative is a general-equilibrium approach. Partial equilibrium describes situations in which either the problem under study has few and small enough connections to other parts of the economy that we can safely ignore them, or that the extension to general equilibrium brings sufficient complications that we need to walk before we run and learn what we can on the *assumption* that those interactions are insignificant. One thing that happens in general equilibrium analyses that generally doesn’t in partial equilibrium is that the distribution of income can change as a consequence of some of the changes under study. A change in the income distribution can lead to a change in aggregate demand because the consumption patterns of major groups of individuals differ. For instance, a shift in income from labor to owners of capital might precipitate a shift in demand from basic foods to “luxuries”

or from consumption to saving. We are getting a bit ahead of ourselves now, because we have yet to introduce the study of demand, but we believe there is sufficient intuition about what “demand” and “consumption” involve for the reader to take away a satisfactory preview impression of what to expect in the general equilibrium analysis of income distribution.

1.12 Production Functions in Achaemenid Babylonia

Matthew Stolper’s analysis of tablets from the Murašû Archive (Stolper 1985, Chapter VI) contains a number of tables showing various inputs into agricultural operations and some indicators of outputs: numbers of oxen or cows; equipment such as plows and harness; rental prices, in terms of grain, of these variable inputs and land, and the apparent rent on land; outputs of barley; influence of a plot’s location adjacent to a canal, which of course offers a more convenient supply of water as well as the possibility of transportation of harvested crop. He infers the considerable value of location next to a canal, but otherwise a generally low price of land relative to the costs of the moveable inputs. These are the classic ingredients of a production function, but analyzed without the benefit of the production function as an organizing concept. At the risk of using an interesting and excellent piece of work as a negative example, Stolper relates outputs to the quantity of a single input at a time, which doesn’t take advantage of the information on how the presence of one input affects the productivity of another – with the exception of the water in the canals, which he doesn’t really acknowledge as another input. Also the production function framework for thinking about everyday work offers an adding-up discipline that is useful – it helps the student account for everything that goes into the production and relate those things to everything that comes out. In a particularly interesting subset of these texts, Stolper ran into this adding up issue and intuitively recognized it but did not appreciate the full implications of his conclusions. We turn to this case.

Reinforcing, in his judgment, the conclusion of typically low land prices are four texts

recording what Stolper calls “agreements to cultivate land in partnership” (130). It could be called a share-rental agreement, such as is common throughout both the developing and industrialized world today. These four texts describe agreements (contracts) between owners and renters of land, and Stolper interprets the agreement in such a way that the land owner furnished his land, both parties supplied animals, equipment, laborers, and so on, in equal quantities, and then they shared the crop equally. Stolper noticed something odd about this arrangement – that it didn’t leave any return for the people supplying the land – but did not pursue the matter other than to interpret the case as further evidence of cheap land.

With the application of some simple production theory, it’s easy to show that this interpretation of the tablets implies that the land owners were letting the renters use their land rent-free. Some contemporary scholars would be inclined to favor such an interpretation if they actually worked it out themselves, but the issue of “free” land outside a distant frontier region, which this wasn’t, raises more questions than it answers satisfactorily. An alternative interpretation of these results is that the tablet evidence was incomplete but was translated as if it were complete.

If the land owner supplies the land and half of everything else and the other fellow supplies half of everything else, how do you decide what the income share of land, labor, equipment, and so forth, are? If 50% goes to half of the labor input (the “other fellow”) and 50% goes to half the labor income and all the rental (land) income, what are the shares of labor and land in production? If you’ve looked at these numbers and thought that something was funny, you’re right. Follow this: Let s_N be the share of labor’s contribution to the output and s_L be land’s share (the “share” concepts from production theory – they refer to the share of the output produced by the specified inputs; the shares will add up to 1). Start with the “other fellow”: he gets half of labor’s share of output and that’s it; that is equal to half of the total product. In other words, $0.5s_N R = 0.5R$. Now, the land owner gets the other half of the output, while his contributions are half of the labor and all of the land. So he has a claim on half of labor’s contribution to output plus all of land’s

contribution to output. Expressed as an equation, that is: $0.5s_N R + s_L R = 0.5R$. Now, from the other fellow's equation we get the solution that $s_N = 1.0$; plug that into the land owner's equation and we get the result that $s_L = 0$. We could set this up as a set of two simultaneous equations (s_N and s_L are the variables) and use matrix algebra to get numerical solutions for s_N and s_L , and we also get $s_N = 1$ and $s_L = 0$. The social interpretation of this is that the contribution of land to production was zero – at least if people were able to claim what they had produced. (You may ask: “but how did they know what they, or the inputs they

supplied, produced?” Answer: observation and passing down the information in a social information storage and retrieval system.) Either the land in question was at the absolute spatial edge of economically usable land (along the lines of the von Thünen model; see Chapter 11), leaving zero revenue net of transportation costs for land rent, or the observations are questionable. Actually, one of the observations simply has to be wrong; what's questionable is which one it is – that they split the output down the middle, or that they each supplied half of everything else.

References

- Belfiore, C.M., P.M. Day, A. Hein, V. Kilikoglou, V. La Rosa, P. Mazzoleni, and A. Pezzino. 2007. “Petrographic and Chemical Characterization of Pottery Production of the Late Minoan I Kiln at Hagia Triada, Crete.” *Achaeometry* 49: 621–653.
- Boni, M., G. Di Maio, R. Frei, and M. Villa. 2000. “Lead Isotopic Evidence for a Mixed Provenance for Roman Water Pipes from Pompeii.” *Archaeometry* 42: 201–208.
- Chenery, Hollis B. 1948. “Engineering Production Functions.” *Quarterly Journal of Economics* 63: 507–531.
- Fishlow, Albert. 1976. “Brazilian Size Distribution of Income.” In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 59–75.
- Forbes, R.J. 1954. “Chemical, Culinary, and Cosmetic Arts.” In *A History of Technology*, Vol. I. *From Early Times to the Fall of Ancient Empires*, edited by Charles Singer, E.J. Holmyard, and A.R. Hall. New York: Oxford University Press, pp. 238–298.
- Forster, E.S., and Edward H. Heffner, translators. 1955. *Lucius Junius Moderatus Columella on Agriculture and Trees*, Vol. III. *Res Rustica X–XII, De Arboribus*. Loeb Classical Library. Cambridge MA: Harvard University Press.
- Freestone, Ian C. 1995. “Ceramic Petrography.” *American Journal of Archaeology* 99: 111–115.
- Frier, Bruce Woodward. 1977. “The Rental Market in Early Imperial Rome,” *Journal of Roman Studies* 67: 27–37.
- Hein, A., and H. Mommsen. 1999. “Element Concentration Distributions and Most Discriminating Elements for Provenancing by Neutron Activation Analyses of Ceramics from Bronze Age Sites in Greece.” *Journal of Archaeological Science* 26: 1053–1058.
- Jain, Shail. 1975. *Size Distribution of Income; A Compilation of Data*. Washington, D.C.: International Bank for Reconstruction and Development.
- Laurence, Ray. 1999. *The Roads of Roman Italy*. London: Routledge.
- Lucas, A., and J.R. Harris. 1962. *Ancient Egyptian Materials and Industries*, 4th edn. London: Edward Arnold.
- Marsden, James, David Pingry, and Andrew Whinston. 1974. “Engineering Foundations of Production Functions.” *Journal of Economic Theory* 9: 124–140.
- Moorey, P.R.S. 1994. *Ancient Mesopotamian Materials and Industries; The Archaeological Evidence*. Oxford: Clarendon Press.
- Quinn, P.S., and P.M. Day. 2007. “Calcareous Microfossils in Bronze Age Aegean Ceramics: Illuminating Technology and Provenance.” *Archaeometry* 49: 775–793.
- Rackham, H., translator. 1950. *Pliny, Natural History*, Vol. V. *Libri XVII–XIX*. Loeb Classical Library. Cambridge MA: Harvard University Press.
- Smith, Vernon L. 1961. *Investment and Production; A Study in the Theory of the Capital-Using Enterprise*. Cambridge MA: Harvard University Press.
- Stambaugh, John E. 1988. *The Ancient Roman City*. Baltimore MD: Johns Hopkins University Press.
- Stolper, Matthew W. 1985. *Entrepreneurs and Empire; The Murašû Archive, the Murašû Firm, and Persian Rule in Babylonia*. Istanbul: Nederlands Historisch-Archaeologisch Instituut te Istanbul.
- Stos, Zofia A., and Noel H. Gale. 2006. “Lead Isotope and Chemical Analysis of Slags from Chrysokamino.” In *Hesperia Supplements* 36, *The Chrysokamino Metallurgy Workshop and its Territory*, edited by P.P. Betancourt. Princeton NJ: American School of Classical Studies at Athens, pp. 229–319.

- Stos-Gale, Zofia. 2001. "Minoan Foreign Relations and Copper Metallurgy in MM III-LM III Crete." In *The Social Context of Technological Change: Egypt and the Near East, 1650–1550*, edited by A.J. Shortland. Oxford: Oxbow Books, pp. 195–210.
- Vandiver, Pamela, and Charles S. Tumosa. 1995. "Xeroradiographic Imaging." *American Journal of Archaeology* 99: 121–124.
- Vaughan, Sarah J. 1995. "Ceramic Petrology and Petrography in the Aegean." *American Journal of Archaeology* 99: 115–117.
- Webb, Richard C. 1976. "The Distribution of Income in Peru." In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 11–25.
- Weisskoff, Richard. 1976. "Income Distribution and Economic Growth in Puerto Rico, Argentina and Mexico." In *Income Distribution in Latin America*, edited by Alejandro Foxley. Cambridge: Cambridge University Press, pp. 27–58.

Suggested Readings

- Beattie, Bruce R., and C. Robert Taylor. 1985. *The Economics of Production*. New York: John Wiley & Sons, Inc.
- Becker, Gary S. 1971. *Economic Theory*. New York: Knopf. Chapters 7–8.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapter 6.
- Varian, Hal R. 1999. *Intermediate Microeconomics; A Modern Approach* ["Baby Varian"], 5th edn. New York: Norton. Chapter 18.
- Varian, Hal R. 1992. *Microeconomic Analysis*, 3rd edn. New York: Norton. Chapter 1.

Notes

- Section 1.5 shows the full expressions of some popular production functions. These formulas are called "functional forms."
- Lucas and Harris (1962, 150–154) on dyeing; Forbes (1954, 249–250) on dyeing; Moorey (1994, 144–150 on firing pottery, 240–301 on base metals). Moorey (1994, 150) notes the fact that kilns permit more efficient use of fuel; translated into the language of the production function, there is a tradeoff between the use of capital embodied in a kiln and the amount of fuel used in the production of a given quantity of pottery. Lucas and Harris (1962, 371–372) hint at the capital–fuel tradeoff as Egyptian potters moved from simply covering the pots to be baked with a heap of animal dung in Pre-Dynastic times and using straw, chaff, reeds, and so forth, for fuel, to surrounding the heap with a low, clay wall and the dung covering replaced by clay, to finally a true kiln, the use of which must have been well established by the Early Kingdom.
- A sample of metals analyses: Boni *et al.* 2000; Stos-Gale 2001; Stos and Gale 2006. A sample of ceramic analyses: Freestone 1995; Vaughan, 1995; Vandiver and Tumosa 1995; Hein and Mommsen 1999; Belfiore *et al.* 2007; Quinn and Day 2007.
- For instance, the production function Smith (1961, 44) derived for the multiple-pass regeneration process (use of fuller's earth as a catalyst in purifying vegetable oil) is $y = Ax_1[1 - Br^{\gamma X_2/x_1}]$, where $A = \alpha/1-r$, $B = r^{1-[\theta_f/(\theta_r + \theta_f)]}$, and $\gamma = H/(\theta_f + \theta_r)$.

The output level of purified oil is y , x_1 is the quantity of fuller's earth, and X_2 is the capacity of the adsorptive equipment (the capital). The definitions of the engineering parameters are: r is the capacity of the adsorbent after its regeneration relative to before; θ_f is the hours per pass in the filtering phase of the process and θ_r is the regeneration time per pass; H is the hours per year of operation; α is a proportional constant representing the ratio of output in the initial pass to the size of the adsorbent charge in that pass; and β is the corresponding ratio of the adsorbent facility to the initial adsorbent charge. The specialization of the engineering production function to the process (or product) it describes is responsible for this complication. While the Cobb–Douglas, CES, and translog functions can be used to approximate this process as well as innumerable others, the engineering production function can yield insights into the choice behavior toward one specific process or product, contingent upon the technology. Whether the engineering information available on many ancient production processes, such as various metallurgical operations, is sufficient to develop engineering production functions along these lines or not is less important than the alternative perspective on ancient production behavior that this concept opens. The production function focuses our attention on the choices available to ancient producers, within the confines of the

technologies they used, somewhat expanding our horizons beyond the relatively rigid combinations of materials and time incidentally implied by strict readings of many of the physical science studies of these ancient technologies.

- 5 I say “consumption-oriented” because the production of housing is implicit in the example, as well as consumption. We deal with some of the peculiarities of housing as a good in Chapter 12.
- 6 Stambaugh (1988, Chapter 8) identifies the public services offered in contemporary (implicitly United

States) and ancient Roman cities and considers a number of reasons for the absences of public (or even private, sometimes) provision of some of them in the ancient cities. He explicitly declines to apologize for what he believes some readers may consider the use of “modernizing” concepts.

- 7 This is a matter of the “derived demand” for roads – a demand derived from people who want to travel and carry things. We’ll introduce derived demand in Chapter 2.

Cost and Supply

The theory of cost is closely related to the theory of production. Intuitively, “productivity” is pretty much the inverse of “costliness”: the more productive one is, the less costly. Production theory describes how much one can produce with a given set of resources, as well as how the resource requirements increase as production increases. Its focus is on quantities of things. The theory of cost focuses on the sacrifices required to produce the quantities of things that production theory describes. They are two aspects of the same phenomenon. We may think of the cost of 200 bales of cotton in terms of money: say, 2 shekels per bale, times 200, that’s 400 shekels. Alternatively, we could think in terms of the wage (labor) payments and land rentals we incurred: say, 10 person-years of labor services at 32 shekels per year and 10 hectares at 8 shekels per year rental, which comes to 400 shekels. Yet another view of the sacrifice incurred to obtain the 200 bales of cotton is to notice that the same 10 person-years and 10 hectares of land could have been used to produce 100 bushels of beans. So we’ve “paid” one bushel of beans for every two bales of cotton we’ve consumed. This is the opportunity cost of cotton. The cotton is the opportunity cost of the beans. They’re reciprocals of each other.

As much as the impermanence of many of the material accoutrements of life has reduced the

material remains of antiquity, quantities of things still have greater likelihood of survival than does material evidence of their costs. Nonetheless, their costs certainly were of first-order importance to the people who consumed and otherwise used them several millennia ago. The cost function is the value equivalent of the materially oriented production function and serves the corresponding purpose of relating input costs to output costs (sections 2.1 and 2.3). The substitution we introduced with production actually takes real time, and the cost approach deals with this temporality with the concept of the length of “the run” – the long run and the short run (section 2.2). The equivalence of quantity (production) and value (cost) approaches to productive activity establishes a number of useful relationships between cost and production that help us make inferences about one or the other of these two concepts even when we can’t observe directly them directly (section 2.3). This is called the primal-dual relationship, quantities being the “primal” and costs the “dual.”

When thinking of costs, it is useful to think carefully about what producers’ objectives may be. Again, motivated by the belief that “leaving stuff on the table” doesn’t characterize human behavior very well, the analytical concept of optimization helps organize thinking about how

producers deal with costs. Optimization can take the form of either maximizing something, such as profit, or minimizing something else, such as cost. The two are equivalent, via the duality of production and cost functions (section 2.4). Some readers may worry that optimizing, especially with regard to the concept of profit, is anachronistic for the ancient world.¹ That is, in fact, a misconception, but one to whose merits or demerits we will not devote much direct attention, leaving both the analytical and empirical counterparts to that process in the ancient world as proof to be found in the present pudding: if the many discussions of the present text do not convince the reader of the usefulness to the contemporary scholar of the concept of optimization as an analytical device for understanding ancient behavior, no amount of pious exhortation in the abstract will be convincing either. Give it a try and see what you think – but just give it a try before you see what you think.

The matter of product supply is closely related to both cost and production, but is distinguishable from both (section 2.5). How much of various inputs that producers want to use is related to the benefits and costs of using a bit more or a bit less of each (section 2.6). The benefits derive from the goods or services produced and the costs from how much the inputs could produce somewhere else (section 2.7), and the comparison implicit in the juxtaposition of costs and benefits suggests some predictability in the quantities of inputs used (section 2.8). The concept of the firm as a group of people who plan and organize the activities involved in any line of production is generalizable as easily to the ancient farm family as to the modern corporation – probably more easily, in fact, because of the smaller scale (section 2.9). Section 2.10 returns to the subject of the cost function with a little more explicitness than we gave it in section 2.1.

The third-from-last section of the chapter (section 2.11) applies the chapter's concepts to a specific case of a subject of wide interest in ancient studies – pottery. Taking as a reference point a recent publication on the value of Mycenaean pottery at Ugarit, this section is the first

of a series of sections throughout the next few chapters applying basic economic concepts to an important ancient artifact type. We used pottery production as an example in Chapter 1, and this section develops the cost side of the same subject. The subsequent treatments of this topic – in demand and competition theory in Chapters 3 and 4 – focus more directly on the subject of interest to the article's author, but the full array of perspectives together demonstrate the organization of thinking about pottery that contemporary economics can offer. The penultimate section (section 2.12) discusses an archaeological-economic analysis of some changes in Minoan pottery between the end of the Old Palace Period and the beginning of the New Palace Period. The final section of the chapter (section 2.13) moves from production at the scale of a single producer to the scale of an entire economy, where an entirely different set of constraints appears – constraints that no individual producer is likely to see but that affect them all with an iron force.

Many of the cost and price data from the ancient world, upon close study, have been found quite difficult to use. Frequently it is difficult to identify precisely the quantity of an input or a good to which a cost or price is attached, much less its original quality.² Those data often convey cloudy and uncertain information. One potential scholarly reaction to the state of these data is, "With these unusable data, who needs theory?" That's just one more reason to have a reasonable command of theory! Offering explanations on the basis of economic data such as these without the conceptual guidance of theory is comparable to eschewing a flashlight because it's dark outside. The conceptual understanding of what variables tend to co-vary, and which one moves which way in response to movement in the other can be worth a considerable volume of opaque data. It's worth recording that empirical researchers using contemporary data never, ever, have data that fit their purposes and needs perfectly, either – often not even satisfactorily. Compromises and approximations are the rule. It's not clear that poor data are a qualitatively distinct problem facing

students of the ancient world's economic behavior, and certainly not one that justifies jettisoning the primary body of theory guiding the analysis of economic behavior.

2.1 The Cost Function

A cost function relates the costs of particular quantities of some product to those quantities: $C = f(Q)$, where the functional notation f does not imply that we are using the same function that we did for production. C is the total cost of production, and Q is the total output. Figure 2.1 shows a total cost curve. Unlike the total product curve, the total cost curve (TC) has a positive intercept, indicating that we incur some cost just to start production, even if we produce no units of Q at all. Curve TVC is total variable costs – the costs that we incur once production begins; it does begin at the origin. Figure 2.2 shows three average cost concepts associated with the total cost curve. The average fixed cost, AFC, of Figure 2.2 is simply distance OF from Figure 2.1, divided by each quantity of output of product Q . It always gets smaller as output increases; for sufficiently large output it becomes negligible. Average variable cost, AVC, in Figure 2.2 is constructed by dividing the cost of each point on the TVC curve in Figure 2.1 by the output associated with that cost. Geometrically, this is equivalent to drawing a line from the origin to each point

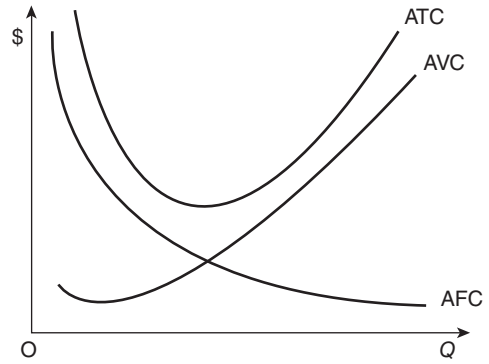


Figure 2.2 Average total, variable, and fixed cost curves.

on the total variable cost curve. The average total cost curve is constructed analogously from the total cost curve.

The marginal cost of producing any unit of output Q is the addition to total cost incurred by producing one more unit. Similarly to the way we showed marginal product in Figure 1.4, Figure 2.3 shows the marginal cost concept as the change in total cost divided by the change in output: $(C_1 - C_0) \div (Q_1 - Q_0)$, as the increment of Q_1 over Q_0 gets very small. Figure 2.4 shows the marginal cost curve (MC), which intersects the average variable cost curve (AVC) from below at the minimum point on the AVC curve. In other words, $MC < AVC$ at all quantities of Q smaller than the Q that gives the smallest average

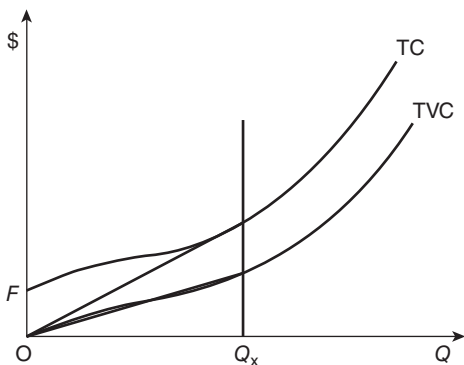


Figure 2.1 Total cost curves: total and variable.

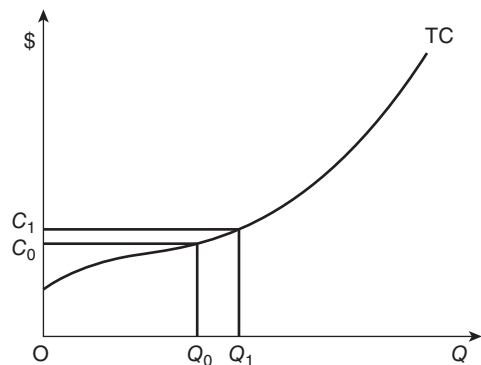


Figure 2.3 The marginal cost concept.

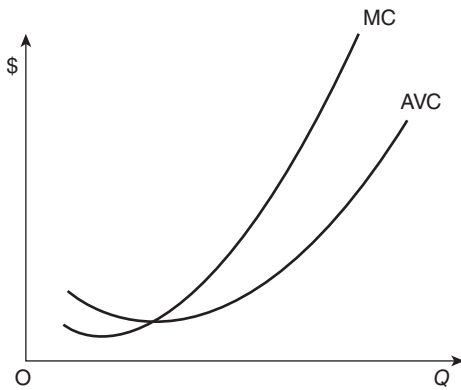


Figure 2.4 Marginal and average variable cost curves.

cost, and $MC > AVC$ at all quantities of Q larger than the average-cost-minimizing level of output. (Why do we derive the marginal cost curve from the total cost curve rather than from the TVC curve? It makes no difference: except for the fixed cost difference represented by the intercept of the TC curve, the TC and TVC curves are the same. Any differences in the levels of the TC curve associated with different levels of output will be exactly the same differences that are associated with different outputs on the TVC curve.) Geometrically, the slope of the tangent to the total cost curve (either TC or TVC) at each point is the marginal cost of the associated level of output Q .

2.2 Short Run and Long Run

The “short” and “long” runs are defined in relation to one another, and it is sensible to discuss the long run first. The “long run” can be thought of as a planning period on the part of the agent undertaking the productive activity. It characterizes the production plans over some extended time period, during which the producer can vary the levels of some inputs that are difficult to vary quickly. For example, it may be possible to build a “factory,” or set of equipment types and a protective facility to keep them from the elements in several discrete increments; it may take time to accumulate the resources to construct

the increments, or the growth in demand for the products coming out of it may take time to materialize. Accordingly, there may be several discrete sizes of “factory” that can be built, each with its own set of short-run total, average, and marginal costs. The producer always operates according to short-run costs.

However, a long-run average cost (LRAC) curve can be drawn as a planning device as a line tangent to each of the short-run average cost (SRAC) curves. The tangencies of the LRAC curve with the SRAC curves will occur at the minimum point on the SRAC curve only for the SRAC curve that corresponds to the minimum point on the LRAC curve. This is a moderately tricky concept, and Figure 2.5 shows the relationship. $SRAC_1$ is tangent with LRAC at the scale of production Q_1 . The marginal cost of production at the scale of “plant” associated with $SRAC_1$ is MC_1 . Production will take place at a level of output somewhat below the short-run average-cost-minimizing level, and the marginal cost actually incurred at plant-scale 1 is identified by the dashed line drawn from the tangency of $SRAC_1$ and LRAC. Similarly for facilities of scales 2 and 3. As drawn, facility-size 4 gives the long-run minimum average cost at output level Q_4 . We have not drawn the long-run marginal cost curve, simply to avoid further clutter on an already busy diagram, but it would be constructed by connecting the marginal cost incurred at each facility size (for example, marginal cost of operating facility 1 is C_1).

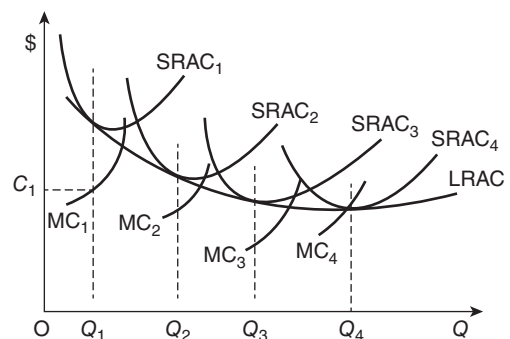


Figure 2.5 Short- and long-run average and variable cost curves.

2.3 The Relationship between Cost and Production

We expressed a very general form of the cost function above as $C = f(Q)$, but we know that the costs of producing Q depend on the factor prices for land and labor (or whatever factors we put in the production function). This means that the cost function can be expressed also as $C = f(w, r)$, where w and r are the wage and rental rates on labor and land. This pair of alternative expressions for the cost function hints broadly at the relationship between the production function and the cost function. Recall the Cobb–Douglas production function: $Q = AN^\alpha L^\beta$. We know that the total cost of producing Q can be expressed as the sum of what we had to pay for the services of the inputs: $C = wN + rL$. From the discussion surrounding Figure 1.6, we also know that the ratio of input prices equals the marginal rate of technical substitution at the efficient combination of inputs. (The MRTS just happens to equal the ratio of marginal products.) The production function itself, plus this condition characterizing the efficient input combination for production, gives us two equations in two variables, which lets us solve for the efficient levels of land (L^*) and labor (N^*) inputs in terms of the wage rate, the rental ratio, and the technical parameters of the production function (α , β , and A). We can substitute these solutions for L and N back into the total cost equation, $C = wN^* + rL^*$, to get the exact functional form of the Cobb–Douglas cost function: $C = A^{-1}(w/\alpha)^\alpha(r/\beta)^\beta Q^{1/(\alpha+\beta)}$.

Under constant returns to scale ($\alpha + \beta = 1$), the unit cost of Q is constant as we produce more Q . If there are decreasing returns to scale ($\alpha + \beta < 1$), there will be increasing costs of producing larger amounts of Q (the exponent on Q will be greater than 1); conversely, with increasing returns to scale, the unit cost of Q will fall as more Q is produced. With a higher wage rate (cost of labor), the total cost will be higher, and conversely for a lower wage rate.

Consider a somewhat closer look at this cost function for the Cobb–Douglas form. Under constant returns to scale, the exponent of Q is 1. Under those circumstances, the marginal cost

function (so far, this is a total cost function) will be just $C = A^{-1}(w/\alpha)^\alpha(r/\beta)^\beta$. In other words, the marginal cost curve (in the graph space of $C-Q$) is flat at a level defined by w and r (and the technical parameters). How do we reconcile this with virtually every diagrammatic depiction of the marginal cost curve as sloping upward? There are two ways. The first is to appeal to some difficulties the agent in charge of production has in coordinating all the activities involved: as the scale of the production operation (one could call it the “firm”) increases, it becomes unwieldy in some ways. The second is to look directly at the factor prices, w and r . Increases in the scale of the production operation could bid up the prices of these inputs. A small production operation (firm) in a large economy (at least large relative to the firm under consideration) is unlikely to have such influence unless the factors are particularly specialized. Third, the cost function derived from the production function addresses long-run cost; the quantities of all inputs are variable. In the short run, some important inputs may be fixed in quantity, which would cause the marginal cost curve to slope upwards sharply. The marginal cost curves we have drawn so far have been primarily short-run curves; connecting the dots in Figure 2.5 would show that the long-run marginal cost curve is much flatter than any of the short-run marginal cost curves. Fourth, an industry is composed of a number of producers of the same or quite similar products. Each firm might try to expand its output, and hence its demand for inputs, at the same time; individually their effects on the prices of factors would be negligible but altogether they could be quite effective.

Conison (2012, 130–143) uses a cost-function model to analyze Roman wine producers’ attitudes toward wine quality and quantity tradeoffs. He concludes that they did not make conscious efforts to modulate the quality of their grapes, but instead focused exclusively on maximizing the quantity of wine their grapes could produce. Quality variations in the wines of the time were attributable solely to the locations of the vineyards, and possibly the competence of the producers.

2.4 Producers' Objectives

So far we have not put the producer in much context. The way we can do that is to specify a goal, or objective, for him or her. The most common structures of this problem are profit-maximization and cost-minimization objectives. The two are closely related to each other, although the profit-maximization goal is more realistic. The conditions for maximizing profit and for minimizing cost are the same: the ratio of input prices must equal the MRTS. We will discuss profit further below but for the moment let's write the profit maximization problem as "maximize $\pi = pQ - C$," where π represents profits (simply the difference between what the producer can get from selling the output and what it costs him to produce it), and C is the cost of production: $C = rL + wN$. The producer is free to choose values of L and N , but can do nothing to affect p , r , or w . Now, recognize that the choices of L and N affect Q as well as C : $\pi = pf(L, N) - rL + wN$. This is a case where knowing some elementary calculus would help considerably, but we can walk through what the producer does in a fashion that replicates the results of the calculus without actually doing it at all.

The producer will look over a range of possible values of L and N separately to see how each choice would affect the value of π . That is, without varying the choices of N , he scans the consequences of different choices of L , and vice versa for scanning choices of N . To see the effect of changing the values of N and L on the revenue term $pf(L, N)$ (this is revenue because it's what he gets from selling his wares), recall from Figure 1.2 and that discussion how the marginal product of an input is found: it's the change in Q divided by the change in the input, where the change in the input gets very small. Then the change in profit from a change in N is $\Delta\pi/\Delta N = p\Delta Q/\Delta N - w\Delta N/\Delta N = p\Delta Q/\Delta N - w$. The symbol Δ just means "change in"; thus ΔN means "change in N , or labor used." Looking over the profit formulation, we notice that the output is multiplied by the product price p , so we need to multiply the marginal product of N by the product price; then the next term in the formulation that contains N is the $-wN$ term, so a change in N will affect that term as the wage times the change in N , or $-w\Delta N$, for every change in N ,

ΔN , which cancels to be simply 1 times the wage rate. Let's suppose the producer contemplates possible values of N in an order from smaller values to larger (that's for our convenience, not necessarily his). Larger values of N will give larger values of Q , although as the values of N get bigger, the increments in Q will get smaller – and they may even become negative, although that's not necessary for our results.³ Now turn to the effect of larger values of N on the $-w$ term: as long as the ΔN increments stay the same size, this term will become more negative at the constant rate of w . So our producer's formulation (he certainly need not see it as a formula) has a positive term that gets bigger at a decreasing rate and a negative term that gets bigger at a constant rate. Sooner or later, the two terms will exactly cancel each other and $\Delta\pi/\Delta N$ will equal zero. That's what we're looking for: $\Delta\pi/\Delta N = 0$. Now, the producer does exactly the same sort of scan for values of L and looks for the same result: $\Delta\pi/\Delta L = p\Delta Q/\Delta L - r = 0$. With some types of formulations, these zero points of additional (or marginal) profit could identify minimum profit, but in this case we are assured that they identify maxima of the responses of profit to changes in these two inputs.

Next recall from our discussion of choosing the optimal combination of inputs for production that we could solve two equations (the production function and one of the marginal product expressions) and get the optimal values of the two inputs. In the present case, we can solve these two "marginal" equations (the formulations for $\Delta\pi/\Delta N$ and $\Delta\pi/\Delta L$) and find the profit maximizing combination of N and L . Let's look at how we could go about doing that, and we will notice several other important characteristics of profit maximization. We can rearrange the formulation for $\Delta\pi/\Delta N$ to get the expression that $p\Delta Q/\Delta N = w$ and the formulation for $\Delta\pi/\Delta L$ so that $p\Delta Q/\Delta L = r$. Recall that $p\Delta Q/\Delta N$ is the value of the marginal product of labor and $p\Delta Q/\Delta L$ is the value of marginal product of land. The ratio of these two values (the p terms cancel) is just the marginal rate of technical substitution. Through some further rearrangements, we can get the result that $(\Delta Q/\Delta L)/(\Delta Q/\Delta N) = r/w$, which is the same lesson we took away from the tangency of the relative price line and the isoquant in Figure 1.2.

There is yet another piece of useful information to be taken away from these marginal profit equations. The pQ term in the original expression is the producer's revenue, so the $p\Delta Q/\Delta N$ and $p\Delta Q/\Delta L$ are marginal revenues. (A marginal revenue, in general, is the change in total revenue contributed by a change in whatever – sales volume, production inputs, and so forth.) The wN and rL terms are costs, so the w and r terms in the marginal equations are marginal costs. For profit maximization, marginal revenue equals marginal cost. The marginal revenues in these equations are partial marginal revenues, in the sense that they are the individual contributions of individual inputs to total revenue, but the lesson generalizes to the case of total revenue. We will see this phenomenon again.

These two marginal profit equations are actually called the “first-order conditions” for optimization – in this instance, a maximization. They assure that the particular values of the “control variables” – labor and land in this case – are associated with no additional change in the value of the “objective function” – π in this case – but they cannot guarantee that this lack of further change in π happens because we’ve “climbed a hill” and reached the top or gone down into a valley and reached the bottom. The zero value of the first order condition tells us that we’ve hit a “flat spot” in the value of our objective function, and we’ve made it either as large or as small as we can make it by continuing to change the value of the control variable. To determine which of these two cases is what we’ve done, there are “second-order conditions,” which tell us whether we’ll get worse – get a smaller value of π – if we keep increasing the value of our control variable (labor or land) or get better – get a larger value of π . Naturally, if things get a bit better when we increase land or labor a bit more, we weren’t at a profit maximum when we hit our “flat spot.” To bring these terms closer to their meanings in calculus, the first-order condition is obtained by differentiating the objective function with respect to the control variables – land and labor in our case. (Differentiating is, roughly speaking, finding out how the value of a function changes in response to very small changes in the independent variables – smaller versions of the Δ s we’ve used above.) This is equivalent to finding the slope of the total cost curve at any point;

setting the derivative expression equal to zero finds a maximum or minimum – the flat spot we found when $\Delta\pi/\Delta L$ was zero. Differentiating the first derivative – that is, taking the second derivatives of the objective function with respect to L and N – tells us how the first derivative itself is changing. If the second derivative is negative, the function value starts falling (getting smaller as the value of the control variable keeps increasing) just after the point where the first derivative was zero, which means we’ve climbed up to the top of a hill (think of the TP curve in Figure 1.4) and now we’d be going down if we continued; the first derivative would actually be negative if we kept increasing the control variable. If the second derivative is positive, we’ve reached the bottom of a trough and have started to climb back out. For some mathematical formulations, we know without actually conducting the second differentiation whether we are dealing with a maximum or a minimum, but the second-order test still yields formulaic results that give us more information about the behavioral and technical characteristics of our extremum (either a maximum or a minimum). I generally won’t delve into these intricacies, but I also believe that it will be useful for the reader to understand that some occasional statements that may seem to be “pulled out of a hat” actually have their basis in some closely governed analytical techniques. Since reference to, and brief explanation of, first-order conditions have appeared even in *The New Yorker* (Cassidy 1998), I believe that discussion of such technicalities is not out of place in a scholarly introduction to the application of economic principles.

2.5 Supply Curves

Cost curves address how much it costs a producer to produce various quantities of products as functions of the amount produced and of the underlying costs of inputs. The supply curve (sometimes called “function” or “schedule”) describes how much output a producer wants to offer at a range of product prices. A supply function has the general form $Q^s = f(p)$, where we use the superscript to distinguish the fact that there is some hypothetical quality to the supply function, because the producer / supplier can

offer a number of quantities of his product that he in fact chooses not to produce. As price p goes up, the supply of product Q goes up. The supply elasticity is $(\Delta Q^s/Q^s) \div (\Delta p/p)$ and is always positive or zero. The supply curve has the implication that most of the points along it are just proposals – firm proposals but still proposals. The point actually chosen on it (the actual quantity supplied) will depend on the purchase offers the producer receives (which takes us into the subject of demand, and we must not get too far ahead of ourselves in the explanations).

Since the quantity of output that producers are willing to offer at any particular product price clearly depends on how much it costs them to produce it, it is logical to look to the production and cost functions for guidance on the supply curve. We can, in fact, derive a supply curve from a production function by applying one of the producer's possible objectives to the problem. What we will do to develop the supply function is pose the producer with the problem of maximizing profit, but we will add a constraint that we did not use before – that the production technology involved in the profit maximization be described by a production function. That is, when maximizing profit by picking quantities of labor and land that give the largest difference between sales revenue and production costs, those values of labor and land are affected by the production technology the producer has to use. Our producer's problem will be: Maximize $\pi = pQ - wN - rL$, subject to $Q = A(N^\alpha + L^\beta)$, in which the parameters α and β both have values greater than 1.⁴ We have not studied a constrained optimization problem yet, although we have introduced optimization and the first- and second-order conditions for finding out if a particular choice is an optimum. The way we will handle this problem is with the method of Lagrange multipliers, in which the constraint function, times a multiplier, for which we can use λ , is added to the objective function. The Lagrangean function is then: $\mathcal{L} = pQ - wN - rL + \lambda[Q - A(N^\alpha + L^\beta)]$. In this constrained function we vary the usual control variables of the economic agent (the inputs labor and land, as well as output in this case) and the multiplier, setting the first derivatives (the $\Delta Q/\Delta L$, $\Delta Q/\Delta N$, and $\Delta Q/\Delta \lambda$) of

each of them to zero for a maximum. This set of operations gives us three separate relationships that we can solve for the profit-maximizing values of land and labor as well as the value of the multiplier λ . This optimal value of λ has the interpretation of the value of changes in the objective function – that is, in this case, the value of expanded output. Then we can substitute the profit-maximizing values of N and L into the expression that maximizes the value of λ , which is just the production function. These substitutions give us an expression for output, Q , in terms of the output price, p , and the two input prices, w and r : $Q = A[(p\alpha/w)^{\alpha/\alpha-1} + (p\beta/r)^{\beta/\beta-1}]$. We are particularly interested in the relationship between Q and p , since the supply curve tells how the offers of Q vary as the output price varies. The output elasticities α and β are both positive and greater than 1, so the amount of Q offered will increase as the output price increases, although the degree to which an increase in p elicits larger Q depends on the production technology as it is characterized by the α and β parameters. The exponents on A indicate that technological progress (the ability to produce more for the same array of inputs or less) will increase the amount of output the producer is able to offer at any price. The α/w and β/r terms indicate that increases in the producer's costs will reduce the amount she is willing to supply at any particular price. Altogether, this function contains information that sounds quite reasonable.

Please pay careful attention to this analytical technique. It is literally the primary workhorse of contemporary economic analysis, and as such we will use it ourselves again and again throughout this text. Although we will not actually “do” the mathematics involved in these problems, the formulation of choice-behavior problems as constrained optimizations is standard in contemporary economics, and a tremendous amount of information emerges from both the first-order and second-order conditions, which we introduced above. Writing these constrained choice problems as formulas may seem a bit foreign, possibly intimidating, at first, but we will “talk our way through them” and find out just how many words these expressions save. Simply setting up choice problems in this fashion

also lets you see the structure of a behavioral problem very compactly, even if you choose not to go through with all the algebraic details of the actual optimization. As Yogi Berra (the famous American athlete – baseball player – of the 1950s and 1960s) was known to recommend, “You can observe a lot just by watching.” A number of optimization techniques exist, specialized to particular types of problem. The one here is widely used for “static” problems – as contrasted with “dynamic” problems – and its usefulness is restricted to what are called “linear” problems, ones in which the choice variables aren’t multiplied by one another. Those problems in which some of the choice variables are multiplied by one another are distinctly more intricate and are handled by varieties of nonlinear programming. A static analysis simply refers to the fact that the formulation of the problem abstracts from the passage of time: everybody knows that the events take time to occur, but the analysts believe that the aspect(s) of the problem in which they are interested can be exposed well without getting explicitly into the time dimension, which can complicate the logic considerably. As a matter of technical structure, the value of a variable in one time period doesn’t depend on the value of some other variable in either previous or succeeding periods in a static formulation of a problem. A dynamic formulation of a problem is characterized by such interdependence of variables across time periods. The method of analyzing the sensitivity of these Lagrangean choice problems to changes in parameters – such as the sensitivity of input demand to changes in product or factor prices – is known as “comparative statics.” This is another name for studying first- and second-order conditions. In dynamic formulations, studied with variational and optimal control techniques, the analysis of such sensitivities of an entire time path of choice variables is called comparative dynamics. The objective function of dynamic formulations is known as the Hamiltonian function (or just “the Hamiltonian”) instead of the Lagrangean. We will deal with static formulations of problems as often as we can in this work. Despite the popular images of “static” versus “dynamic” that have been encouraged by both scholarly and Madison

Avenue product advertisement, no stigma is attached to the formulation of problems as static.

We have not distinguished between a firm’s (or an individual producer’s) supply curve and an industry supply curve in what we have developed so far. Figures 2.6 through 2.8 show how we can add the supply curves of individual producers to get the supply of the industry. Individual firm supply curves appear in Figure 2.6. They have different price intercepts, indicating that some can enter production at lower prices (costs) than others. Firm 1 (represented by curve q_1) faces a higher entry price than does Firm 2 (curve q_2), but can increase its supply at lower cost than can Firm 2. Firm 3 faces an intermediate entry cost but thereafter has a long, relatively flat stretch in which it can expand its production at little increase in unit cost; however at some point something happens to its capacities, and its expansion costs increase sharply. Take another look at the supply equation in the previous paragraph, and you will be struck by the possible causes for these differences in firm supply curves. First of all, we may not be surprised at all to find a situation in which different producers (firms) had different abilities to expand their supplies. But looking at the supply equation, what could contribute to such sharp differences in performance? If every firm is using the same technology (that is, A , α , and β are identical), and they face the same factor prices w and r , what’s causing the differences? Can we in fact draw these different supply curves as we have in Figure 2.6? Yes, we can, but we have to be able to explain

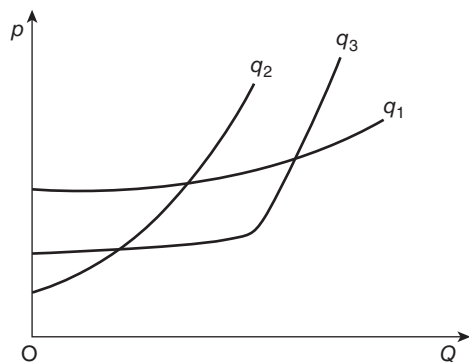


Figure 2.6 Supply curves of individual producers.

why. First, the different producers may in fact be using slightly different technologies. Second, the supply curve we derived above from a production function is a long-run curve, and some of the firms may be stuck in situations where they are unable to adjust the inputs of some of their factors, a consideration that the long-run form presented above does not include. Third, we have abstracted considerably in our specification of the underlying production function when we have said that the only inputs in it are land and labor; there might be a number of other factors used in small, or even not-so-small, proportions that could account for the differences in the supply curves drawn in Figure 2.6. Reference to the basic model (theory) helps us understand and describe the difference between what we observe (in the diagram, but also in the observational world – frequently called the “real” world by people who believe that thinking is not real) and our conceptualizations of it. The model is not incorrect in any meaningful sense of the term, but it may be incomplete – deliberately so because we were focusing on a small number of critical relationships.

Figure 2.7 shows the horizontal summation of the supply curves of the three firms shown in Figure 2.6. At any price p on the ordinate, we can read off the abscissa the quantity supplied by each of our firms. These offerings are independent of one another, so we can just add them – horizontally on the diagram, not vertically. Figure 2.7 represents a case of no interactions at all among these suppliers that would

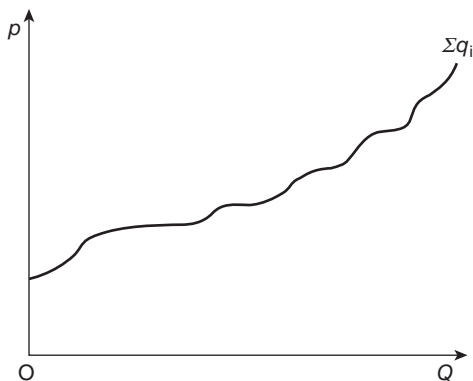


Figure 2.7 Supply curves of individual producers summed horizontally.

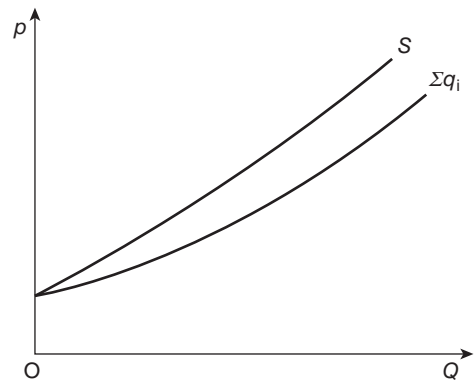


Figure 2.8 An industry supply curve that diverges from the sum of individual producers' supply curves.

feed back onto them to cause any of them to be able to supply less (or more) of their output than they would anticipate being able to supply. But Figure 2.8 represents a case in which there is some interference among the individual firms. As they all try to supply output along their individual supply curves, they are clearly doing something to raise their costs. A likely culprit is that, in trying to expand the scale of their operations, they are bidding up the prices of some of their inputs. Another look back at the supply function will show that this will depress the supply offered at any price, which would account for the difference between the S and Σq_i curves (where the symbol Σ stands for “summation of” – for example, $q_1 + q_2 + q_3 + \dots$).

Before departing the subject of supply curves, we wish to leave several summary messages. First, whenever using the supply curve construct or examining someone else's use of it, clarify for yourself whether you are dealing with a firm's (or individual producer's) supply curve or an industry supply curve. Second, keep in mind several “limiting-case” shapes of supply curves: perfectly elastic and perfectly inelastic, the latter representing the case called “fixed supply.” Figure 2.9 shows the perfectly elastic, or “flat,” supply curve. This shape says that the firm or industry in question is willing to supply any quantity of output Q at the price p^* . Below that price, it will supply nothing; above that price, it will be equally happy to supply any number of units of Q , but p^* is its lower limit. The supply elasticity of this shape is $+\infty$. When a firm

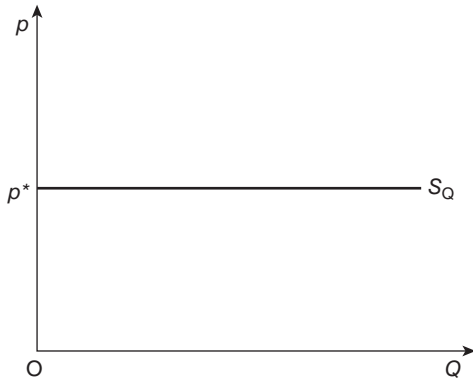


Figure 2.9 Perfectly elastic supply curves.

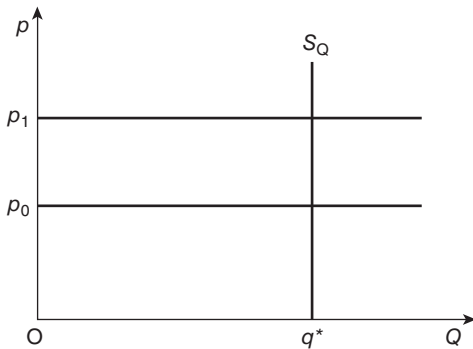


Figure 2.10 Perfectly inelastic supply curve.

has this shape of supply curve, we know that its production function has constant returns to scale. If an industry has this shape supply curve, it is what is called a perfectly competitive industry (concept discussed in a subsequent chapter). Figure 2.10 shows a perfectly inelastic supply curve. The supplier can supply the quantity q^* regardless of the price. Increasing the price from p_0 to p_1 , or to any price for that matter, will not increase the supply. Such a situation clearly represents some underlying fixed inputs. The reader will note that this supply curve, as drawn, indicates that even a zero price will elicit supply q^* , which seems odd. It is odd, and a more likely case of such an absolutely fixed range of a supply curve is shown in Figure 2.11, where the supply curve is perfectly elastic at price p^* from the origin to quantity q^* , at which quantity it is perfectly inelastic. The supply curve in Figure 2.10 is more likely to represent the supply

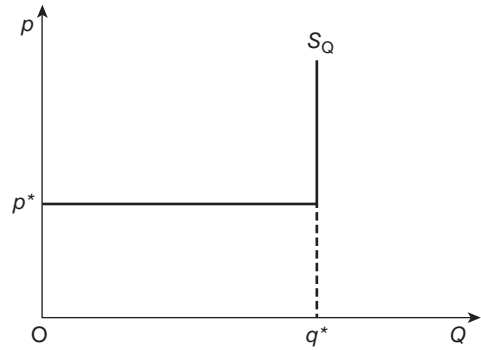


Figure 2.11 L-shaped supply curve: perfectly elastic, then perfectly inelastic.

of a factor, like land (but probably not labor, since labor would have a “reservation price” below which it would not offer itself – a reservation price for labor being based as much on the value of leisure as on the minimum price – in terms of, say, food – required to feed itself). The L-shaped supply curve of Figure 2.11 is likely to characterize the supply response of a firm with a CRS production function and an absolutely fixed factor for which no substitutes are available. Third, consider the shape of the supply curve in Figure 2.12 and think about why it is impossible. Such a downward-sloping shape could characterize a *portion* of an average or marginal cost curve for a firm, but that curve has to turn upward at some point, and producers will operate only on the upward-sloping segments of their cost curves (some monopolistic producers are an exception to this rule, and we will see the downward-sloping

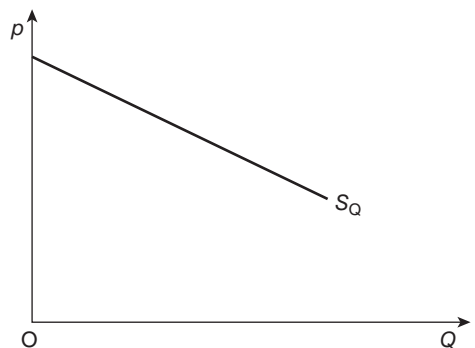


Figure 2.12 Downward-sloping supply curve (impossible).

segment of the cost curve again, although not as a supply curve).

2.6 Demands for Factors of Production

We have not introduced the concept of demand yet. It is common to think in terms of a consumer's demand for various products, but the same concept applies to producers' demands for inputs. The demand for a factor of production identifies equally (i) the quantity of that factor a producer will want to use at a given factor price and (ii) the maximum unit price the producer would be willing to pay for the services of so many units of the factor. In consumer theory, as we will see later, the maximum amount a consumer is willing to pay for a unit of a product to consume is determined by how much satisfaction he gets out of it. In the case of the demand for production inputs, how much the producer is willing to pay is determined in large part by how productive the input is – that is, by its marginal productivity. In fact, the equation for the demand for a factor of production is simply the expression for its value of marginal product. In the case of the Cobb–Douglas production function with two factors (which we did *not* use in derivation of the supply curve), the marginal product equation for labor is $N = p\alpha Q/w$; if we want to get rid of the Q in that expression, we can write the same demand function as $N = (p\alpha AL^\beta/w)^{1/1-\alpha}$, which amounts to exactly the same thing but lets us see quite clearly the roles of the technology parameter A and the quantities of the other factors (only L in this case) in affecting the demand for labor. As the price of the product increases, the demand for labor increases. If the price of labor (w) increases, the quantity of labor demanded falls. With unchanged w , as the quantity of the other factor increases, the demand for labor rises (this wouldn't necessarily be true if we allowed w to change at the same time). If the technology has a larger output elasticity of labor (α), demand for labor is greater. If the overall productivity of the technology, as represented by A , increases, labor demand increases. From the first version of the labor demand equation,

as the output increases, the quantity of labor demanded rises. We can express this quite specific labor demand function in a more general form as $N^d = f(w/p; Q, T)$, where T represents characteristics of the production technology. Note that in the general form of the function, we have made the independent variable w/p rather than simply w . What is important to the demand for labor is the cost of its services relative to the output price. If both w and p change by the same percentage, the quantity of labor demanded will be unchanged. This is an important feature of demand functions.

We should make an important distinction here between “labor demand” and “the quantity of labor demanded.” The term “labor demand” refers to the function. Changes in the independent variables composing the function can change the entire schedule of labor demands – that is, as in Figure 2.13, they can change the position of the demand curve. Along any given demand curve, the “quantity of labor demanded” changes continuously as the relative price of labor services (w/p : that is, price of labor services relative to the product price) changes. Changes in w , p , or w/p together precipitate movements *along* a labor demand curve. Changes in technology parameters or Q cause *shifts* in the curve. A shift in the curve may or may not lead to a change in the quantity of labor demanded; that depends on what, if anything, happens to the supply curve of labor, which we have not discussed yet. So in referring to events in the market for labor we need to distinguish between changes in the quantity of

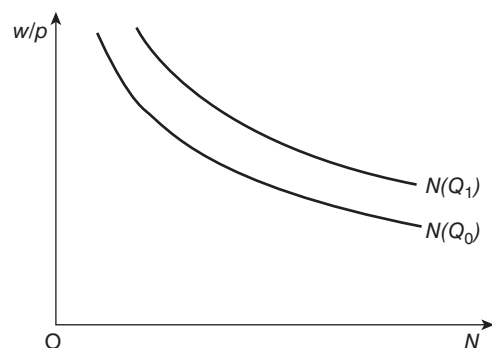


Figure 2.13 An input demand curve.

labor demanded and changes in “labor demand.” They refer to different concepts.

A somewhat more detailed observation is based on the expression for the demand for labor that has the expression for the quantity of the other factor, L^β , in it. The optimally determined value of L will depend on its own service price, r . Consequently r , or r/p , is implicitly in the demand function for labor. Then we can express the function for labor demand somewhat more generally as $N^d = f(w/p, r/p; Q, T)$ in which N^d is negatively related to w/p and positively related to r/p (recall that with only two inputs, the inputs have to be substitutes for each other). In the yet more general case in which there are many inputs in the production function, the demand function for any input depends on the prices of all the inputs: $X_i^d = f(p_1/p_Q, p_2/p_Q, \dots, p_n/p_Q; Q, T)$, where X_i^d is the demand for any input i . The demand for input i is always negatively related to its own price, but it may be positively or negatively related to the price of any other input, depending on whether the two factors are substitutes or complements. It's possible for all other factors to be substitutes for any given factor, but it's not possible for them all to be complements.

The demand for any factor of production is a “derived demand” – derived from the ultimate demand for the product produced with it. This characteristic is reflected in the relationship between the demand for any input and its own price:⁵ in the case of labor demand, the quantity of labor demanded falls with a rise in the relative price w/p . The rise in w/p can come just as well from a fall in p as from a rise in w . Conversely, a rise in the product price will enter the labor demand function as a decrease in w/p , which will expand the demand for labor. What can cause the increase in the product price? The answer, of course, is “an increase in demand for the product,” a topic we have not yet treated directly.

One of the key, descriptive parameters of demand is the elasticity of demand, a deceptively simple term by itself. There are in fact many demand elasticities associated with any input: its “own-price” elasticity, or the percentage change in the quantity of labor demanded in response to a 1% change in w/p ; and $n-1$ “cross-price” elasticities, where n is the number of factors of

production. There are several useful expressions for, and rules about, the elasticities of factor demands, which account for the interfactor relationships imposed by the necessity for total factor costs to add up to the total product cost. In the two-factor case, the own-price elasticity of any factor is $\varepsilon_{11} = -[\theta_1\eta + (1-\theta_1)\sigma] < 0$, where θ_1 is the cost share of factor 1 prior to the price change, $\eta < 0$ is the demand elasticity of the product itself, and σ is the elasticity of substitution (defined positively) between the two factors. The demand elasticity of the other factor with respect to a change in the first factor's price is $\varepsilon_{21} = \theta_1(\sigma - \eta) \gtrless 0$, depending on whether the elasticity of substitution exceeds, equals, or falls short of the elasticity of demand for the product. In the multifactor case, the own-price elasticity is $\theta_i(\sigma_{ii} - \eta)$, and the cross-price elasticity is $\varepsilon_{ij} = \theta_i(\sigma_{ij} - \eta)$, where σ_{ij} is positive for substitutes and negative for complements. The following four rules of derived demand can be inferred from these expressions. (i) The elasticity of demand for a factor will be greater the larger is the elasticity of substitution in production. (ii) The elasticity of demand for a factor is greater the more elastic is the demand for the final product. (iii) The elasticity of demand for a factor is greater, the larger is its share in total cost, in the multifactor case; in the two-factor case, this will be true as long as the elasticity of demand for the final product is greater than the elasticity of substitution. This condition is known as “the importance of being unimportant.” (iv) Not shown in these expressions, but equally general, the elasticity of demand for a factor will be greater, the greater is the elasticity of supply of the other factors.

2.7 Factor Costs in General: Wages and Rents

It is common to think of labor as being paid a wage and land a rent. This distinction goes back to the period in the development of economics when land was considered the original and indestructible factor. You couldn't get rid of any of it – it was there in fixed supply whether you used it or not, so you didn't actually *have* to pay it anything

to elicit its supply. It just soaked up the difference between total product and the marginal product of labor times the units of labor used. The term “rent” was used to designate this characteristic of the payment to land: the payment was unnecessary to draw the factor into production.

Some positive payment, on the other hand, is required to draw mobile labor from one activity to another, or from leisure into labor. If a worker is already working in one activity, it will take at least as much as he gets paid in that activity to draw him into another activity (abstracting from any differences in drudgery or danger between occupations or skills involved in them). Any payment above what he could make in the first activity is not an opportunity cost to the worker – he couldn’t make that additional amount elsewhere. Viewed alternatively, the amount of product that must be sacrificed elsewhere in the economy to secure the laborer’s production in a new occupation is the opportunity cost of the new activity. The net addition to total output is what can be produced in the new activity minus what has to be given up to get the worker into the new activity. When such a difference is zero, workers are said to be paid exactly their opportunity cost. In fact, the difference in payment need not be given to her to attract her to the new activity. If such a difference between factor payment and opportunity cost does exist, it is a rent.

Thus, the payments that are colloquially thought of as “wages” may contain some element of rent. Since we know that the supply of land responds to the demand for it, some portion of what is colloquially called “rent” contains what can be called a “wage,” or opportunity cost; the opportunity cost of land at the edge of expanding cities is what it can produce in agriculture. In general, the part of a factor payment that is determined strictly by the supply conditions of the factor can be called the “wage,” and the portion that is determined by demand, and in fact need not be paid to attract the factor to the occupation under consideration, is a “rent.” We can show this graphically in Figure 2.14. Curve N^S is the supply curve of labor to some activity, and w is the prevailing factor payment to labor. The horizontal line labeled w is in fact a labor demand curve, which just happens to be flat: the quantity of labor that would be demanded at that wage rate is sufficiently larger than the

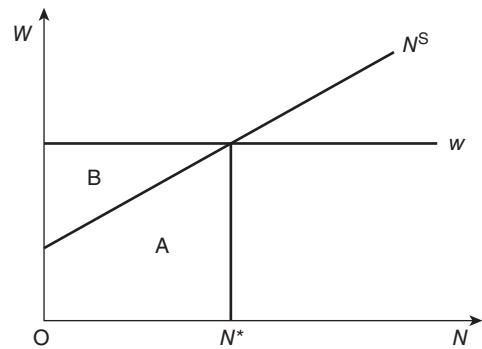


Figure 2.14 Supply and demand curves for an input.

scale of our horizontal axis that it appears that an “unlimited” quantity of labor would be hired at that wage. Labor itself is pickier about how it spends its time. Its supply curve slopes upward, and the intersection of its supply curve with the labor demand curve determines how much labor is actually contracted: N^* is both the quantity of labor demanded and supplied at wage w . (We will discuss in greater detail the equalization of supply and demand curves, for both factors and products, below.) The area under the labor supply curve, from the origin of the graph to the quantity of labor N^* , is the opportunity cost of labor. We can think of the points from O to N^* as individual workers or labor hours. The upward slope of the labor supply curve indicates that the first units supplied are the “easiest” to acquire. Either people can be pried out of leisure at a lower price when they are enjoying lots of leisure, or labor that is less productive elsewhere in the economy can be found at lower prices. As we work our way out the supply curve, we are having to take labor away from more productive alternative uses – either we are having to attract people with better alternatives away from those alternatives or we’re having to get people out of leisure who really like their leisure. There’s no way to tell in general what the composition of the people toward the right end of the labor supply curve is – the people with high-productivity alternatives or the slugs who can’t get out of bed in the morning.

Suppose that two emperors conscript soldiers from among the farmers of their respective

empires. (Their empires consist mostly of farmers, so this source of soldiers is not really discriminatory; each may pick up a few city boys too.) One of them pays his draftees “well,” and the other one, being more frugal, decides to spare the public coffers and pays his draftees only half what the other emperor pays his. Which emperor faces the lower cost of raising his army? Before answering, here is some more information on the two empires: they are both largely agricultural, both use the same technology in their agriculture and both have about the same overall population densities and percentages of their populations in cities. Now, the answer is: they both face pretty much the same cost per soldier. The cost of the soldiers is their opportunity cost, which is the foregone agricultural production that each emperor must sacrifice for his kingdom to raise his army. What the emperor decides to pay his men as soldier’s wages is more of a financing issue than a cost issue.

2.8 Allocation of Factors across Activities

So far we have devoted our attention to the allocation of factors within a single production activity. The guideline for efficient allocation was that the ratio of the marginal technical rate of substitution to the relative factor price be equalized across all factors; stated alternatively, that the ratio of marginal product to factor price be the same for all factors. These intra-activity efficiency conditions can be extended quite easily across activity lines. Take the condition that the marginal product equal the factor price (a variant of the condition expressed just above), for a factor that we will call simply i : $p_Q \text{MPP}_{iQ} = p_i$, where p_Q is the product price, p_i is the price of factor i (the wage rate if we wanted to identify factor i with labor), and MPP_{iQ} is the marginal physical product of factor i in activity Q . For factor j (j not the same as i) $p_Q \text{MPP}_{jQ} = p_j$, where p_j is the price of factor j ’s services and MPP_{jQ} is the marginal physical product of factor j in activity Q . Now, p_Q is in both of those equations. We can rearrange each equation and set them equal to each other since each is equal to the same quantity: $p_Q = p_i / \text{MPP}_{iQ} = p_j / \text{MPP}_{jQ}$; we could

extend the equalization across all the factors used in this activity. Next, the price of the product, p_Q is related to the marginal cost of production as another efficiency condition: $p_Q = x \text{MC}_Q$; x may equal 1, but it generally will be no less than that. Now we focus on the factor prices, p_i and p_j , which all productive activities face equally. Consider product V , which has price p_V ; its intra-activity efficiency conditions will include similar ratios of marginal products and factor prices: $p_V = p_i / \text{MPP}_{iV} = p_j / \text{MPP}_{jV}$. Rearrange the sets of efficiency conditions once more to isolate the factor price: $p_V \text{MPP}_{iV} = p_i = p_Q \text{MPP}_{iQ}$, and we could continue with this equalization of values of marginal product (output price times marginal physical product is called value of marginal product, or VMP) across all activities.

Thus, the efficiency condition for production throughout our economy is that the value of marginal product of each factor be equal in all its uses, and that those VMPs equal the factor price (or some constant times the factor price). We are not claiming that such a condition always occurs. Many things can happen to interrupt it. Products can be taxed, factors can be taxed differently in different activities, and some factors can be stuck in particular activities for a while (although this reason should not be stretched too far because mobility of other factors can compensate for such factor-specific immobility). Other reasons could be found, but these should suffice to give a flavor.

2.9 Organizing Production: The Firm

A firm is a legal entity organized for the purpose of conducting some production activity or array of activities. It can be owned and operated by a single person, it can be owned by several operating partners, it can be owned by absentees (stock holders), it could be owned and operated by government or operated by private contractors. In both ancient Egypt and ancient Mesopotamia, temples acted at least partly as firms. In two Ur III texts found at Umma (dated to between the twenty-first and twentieth centuries B.C.E.) a pottery operation – effectively a firm, despite the possible intervening factor of state ownership – recorded outputs of over

40 types of ceramic vessels and inputs of labor time and materials, as well as credit-debit balances (Potts 1997, 155–157). Whether they (or their directing personnel) thought of themselves as a firm is a moot point and largely immaterial to the use of the concept of the firm as a contemporary analytical device.

Firms have several tasks before them. Some agent in the firm (the owner-operator or a representative) must ensure that the desired factors come together in the desired quantities and places and at the proper times, which generally involves some forecasting of how much can be obtained from the sale of the products once produced. However, considerable obligations frequently must be incurred to the various factor owners (the suppliers of the factors, who are entitled to the payments to the factors they supply) before the firm's director knows for sure how much can be obtained for the output. This director also is supplying a factor of production, for which he or she will be paid a factor price, but the payments to entrepreneurial services and risk-taking must be thought of slightly differently from the payments to other factor services. The owner of the firm generally has the rights to the residual income from production – the gross revenue (pQ) minus the variable factor costs ($wN + rL + \sum p_i$ in our terminology thus far). It is possible that competition among firms will bid this residual income down to nothing more than the factor price that labor receives, but in circumstances when considerable entrepreneurial skills are involved in the direction of the firm's planning and operations, we can expect a positive residual, which can be called profits. These profits are not a factor price but rather a capital gain that accrues to the owner of the firm. They also are not to be confused with the rate of return to "capital," either in the form of financial capital or physical equipment.

If the firm's director is skillful, she will continue to make profits. If unskillful or unlucky, she will make either losses or nothing at all. In the former case, the firm is likely to dissolve, or at least a new director will be found if the director is not the owner. Competition among firms will also place limits on the ability to make such profits over a long period.

If the director of the firm is not the owner, we have a case of what is known contemporarily as separation of ownership and management. There

is no guarantee that managers and owners will have the same incentives. If the manager doesn't get to keep all the profits he makes, his incentive to make profits is diluted and may be distorted into other items of consumption within the firm – power, attractive subordinates, handsome quarters or office, and so forth. Owners recognize this limitation on the natural coincidence of incentives and generally attempt to create some set of compensating incentives and disincentives, ranging from stock options to execution. Any particular society's legal treatment of firms represents at least part of their recognition of this potential problem and their attempts at solution.

So far we have discussed the production of only one product at a time, but many firms produce multiple products, an organization commonly called horizontal integration. The production of multiple products within one "umbrella" organization leaves many of the single-product conclusions intact but requires some generalization of those results in some instances and offers some insights into new issues. First, it is difficult to ignore the productive effects of fixed factors (inputs whose supplies are fixed in the period under consideration) as we have done in the single-product case. Second, we may be able to allocate units of some inputs to particular products ("this much L goes to product A, and this much L goes to product B, making in total that much L "), but that may not be possible for all inputs. An example of such an input is the well-known case of animal feed producing both mutton and wool; how do we tell how much of the feed went to produce the meat and how much went to produce the wool? Both the short and long answers are, "We can't tell." Third, there may be both technical and economic interdependencies among some of the products of the firm. Fourth, the firm may experience costs of switching production back and forth among different products.

The Isin workshop (2017–1975 B.C.E.), whose archive van de Mieroop (1987) reports, has the appearance of a horizontally integrated firm. It employed between 15 and 20 craftsmen in four crafts – carpenters, leatherworkers, reedworkers, and felters – each working under the supervision of an overseer, with two other officials reported who dealt with the receipt of previously processed inputs from other sources and the distribution

of finished products (some possibly for further finishing elsewhere). Products identified are nonceramic containers, footwear, furniture, weapons, musical instruments, vehicles, doors, mats and covers, treatments to cloth, and a *mélange* of other items. Van de Mierop does not attempt to assess the ownership conditions of the workshop – for example, royal, temple, or private – but he does note that more detailed records are kept on intermediate inputs arriving and finished or quasi-finished products leaving, which lends an impression of the workshop comprising a complete accounting unit.

Vertical integration involves housing various related activities, typically stages of production, within a single firm. Vertical integration can reduce the costs of arms-length, market transactions in instances when information about the products of the different production stages would not necessarily be shared by separate firms (situations of asymmetric information, a subject of Chapter 7). An example which Conison (2012, 76–100) has analyzed incisively with concepts from the “New Institutional Economics”⁶ is the Roman wine industry, which never vertically integrated the stages of production, transportation and distribution into single firms because of Roman law’s lack of entity shielding, which would have given priority to the firms’ creditors over the owners’ personal creditors, thus raising great uncertainty in financing.

In the single-product case, the efficiency conditions for production (factor allocation) were that the marginal rate of technical substitution among the variable inputs (we considered only variable inputs in the single-product case) equal the ratio of factor prices;⁷ and that the value of marginal product of each input must equal its price. We did not worry about the marginal productivity of fixed factors, since we couldn’t do anything about them, and we had to pay for them anyway. When a factor is in fixed supply to the firm (entrepreneurial capacity would be one such input), it may be possible to allocate some portion of it to one product and another portion of it to another product. In such cases, the fixed factor takes on characteristics of variable factors, at least variable within the firm, and the ratio of marginal products of the fixed factor in different product lines must be equalized for productive efficiency.⁸

Sometimes a variable or fixed factor contributes to two product lines simultaneously, and the contributions genuinely cannot be separated. The cause of the inseparability might be something along the lines of the feed-mutton-wool case in which the input causes growth in both outputs, or it could create an “atmosphere” that contributed simultaneously to several product lines – something like morale in a labor force. In either case, the allocation of the factor costs to any product would be completely arbitrary on the producer’s part, even if the revenue shares of the products were substantially different (in other words, assigning the full cost of the “atmosphere”-creating input to the dominant product line just because it seems like that product could “carry” the cost would be inefficient). In the case of “nonallocable” factors, the sum of marginal products in all products should be equated to the factor price. This is equivalent to the single-product condition of equating marginal revenue to marginal factor cost, only the revenue is spread across several products.

Recall how we characterized the general expression of the single-factor production function as $Q = f(L, N)$. In the case of the multi-product firm, we can express each production function as $Q_i = f(L, N, Q_j)$. We know that more of each input L and N will contribute to increases in the output of Q_i , but what about the effect of increases in the other product, Q_j ? In general, the effect of increasing Q_i on the output of Q_i can be positive, negative, or zero. Negative and zero effects probably are most common but the synergies of positive effects, at least between some pairs of products, if not between all pairs, cannot be dismissed as one of the reasons for forming firms in the first place. These effects of one output on another output can work in two principal ways. First, when there are fixed factors, the fixed quantity of which can be individually allocated to separate products, producing more of one product must take away from the maximum amount of the other products to which some of the fixed factor could have been allocated. Second, the inputs can influence one another’s marginal productivity. For example, increasing the output of good A while also increasing the output of good B could increase the marginal product of, say, labor, in producing good A. If good A and good B were to be produced with exactly the same production

technology but in separate firms, this effect of increasing the output of good B would have no effect on the marginal productivity of labor in producing good A. Products whose production affects one another positively in this way can be called complementary, and the effect on cost is called an economy of scope, in contrast to economies of scale, which refers to the effect of firm size rather than the effect of the product line. If they have the opposite effect – that is, they decrease the productivity of factors in other product lines, they can be called competing products, or even substitutes in a very real sense, since one product substitutes for the other in the maximum amount of revenue-producing output the firm can produce. This technical effect across products – via effects on inputs – will also appear in what can be called an economic effect: the supply curve of good A, in the case described above, could be upward sloping in the price of product B as well as in its own price. That is, as the price of good B increased, the amount of good A that the firm would be willing to supply would increase, at any given price of good A, because the *effective* price of its labor input has fallen (its marginal productivity has risen).

The cost of switching between products may involve resetting equipment or relocating labor or other costly actions. These are not costs to be simply “written off,” but actions to be considered in marginal analysis as are variable factor costs. The marginal cost of switching between products should be equalized to the marginal contribution to firm revenue of switching.

2.10 A More General Treatment of Cost Functions

A more complete view of the cost function includes the prices of inputs in the function, as well as the quantity of output produced: $C(w, r, p_z, Q) = \min\{wL + rK + p_z Z | f(K, L, Z) \geq Q\}$, where w is the wage rate, r is the rental rate on equipment, p_z represents an array of the prices of any other inputs used, and the symbol “|” means “given that.” The producer adjusts the levels of the inputs, K , L , and Z , to minimize the total cost of producing a level of output at least as large as Q . This cost represents the least cost at which a producer can produce the quantity of output Q ,

facing input prices w , r , and p_z , and a particular production technology, represented by “ f ”) that lets him or her convert labor, equipment and the other inputs into whatever product Q is. A useful characteristic of the cost function is that, whereas $\Delta C/\Delta Q > 0$ is the marginal cost of a unit of output Q (MC_Q) as we noted above, $\Delta C/\Delta w > 0$ is the demand function for labor (which we could denote L^d), $\Delta C/\Delta r > 0$ is the demand function for capital, and so on. We can push these two relationships a bit further: $\Delta[\Delta C/\Delta Q]/\Delta w = \Delta[\Delta C/\Delta w]/\Delta Q > 0$, which can be written alternatively as $\Delta MC_Q/\Delta w = \Delta L^d/\Delta Q > 0$. This says that the change in marginal cost caused by an increase in the price of labor (or of any input) equals the response in the demand for labor caused by an increase in output.

The cost function is called the “dual” of the production function, which itself is called the “primal” form, since one function implies the other. The production function works with quantities while the cost function works with prices. Let’s take the example of the Cobb–Douglas production function that we introduced in Chapter 1 ($Q = Ak^\alpha L^{1-\alpha}$). In section 2.3, we *told* how to get the production function’s cost function, but we didn’t actually *show* how to do it. Now we’ll *show* how. First, since the cost function takes as given that the producer has already found the cost-minimizing methods of producing according to the production function, she has chosen the quantity of each input that will make its marginal product equal to its marginal cost, which is the corresponding factor price (the wage or rental rate). Thus, the quantity of labor used is the amount that will make the value of the marginal product equal the wage rate. The marginal product of labor in the Cobb–Douglas production function, which we’ll denote as MP_L , is $(1-\alpha)AK^\alpha L^{-\alpha}$, or $(1-\alpha)A(K/L)^\alpha$ or $(1-\alpha)Q/L$, and that for capital is $MP_K = \alpha A(K/L)^{\alpha-1} = p\alpha Q/K$. Then the producer adjusts her employment of labor so that $w = pMP_L = p(1-\alpha)AK^\alpha L^{-\alpha}$, and her use of equipment so that $r = pMP_K = p\alpha Q/K$. Now, solve for the cost-minimizing employments of labor and capital (that is, take L and K over to the left-hand sides of these two expressions): $L = p(1-\alpha)Q/w$ and $K = p\alpha Q/r$. Substitute these expressions for cost-minimizing input levels back into the production function, do the

appropriate cancellations, and take p over to the left-hand side of the expression, and we get $p = (1/A)(r/\alpha)^\alpha(w/1-\alpha)^{1-\alpha}$, where p is both the competitive price of the product and its unit cost (we could call it C/Q or c). Now, look at how the unit cost will change with an increase in the wage rate: if the rental rate, r , increases, p will be higher; if the wage rate, w , increases, p will rise. This is just what the general version said in the previous paragraph.⁹

Another useful dual relationship is the profit function, $\pi(p, W) = \max\{pQ - WI\}$, where p is the product price, W is the expression for the entire array of input prices, and I is the array of input levels. What this expression says is, “the profit function π is a function of product prices p and factor input prices W , and specifically, it’s the maximized difference between what the producer gets from selling his product and the costs of the inputs he uses to produce it.” The producer adjusts his output level and the input levels to maximize profit. This is structurally just like the cost function, only it’s a maximization, rather than a minimization, effort on the part of the producer. We could actually rewrite the profit function to be the difference between the value of whatever level of output is produced minus the cost function of producing that level of output: $\pi(p, W) = \max\{pQ - C(W, Q)\}$. Now, the response of the profit function to a change in the output price gives us the producer’s supply function: $\Delta\pi/\Delta p = Q^s > 0$, where the notation Q^s means the quantity of product Q supplied; and its response to a change in one of the input prices gives his derived demand for that input: $-\Delta\pi/\Delta r = K^d > 0$ for the case of the demand for equipment. We can take this set of profit function relationships a step further to find that the effect of a factor price change on the supply function is the same magnitude as the effect of a change in product price on factor demand, but in the opposite direction: $\Delta Q^s/\Delta W_i$ (where W_i is the price – “wage” – of the i^{th} factor) $= -\Delta K^d/\Delta p < 0$. This is yet another way of showing how we can derive supply curves and input (factor) demand curves – they happen because producers try to maximize profit.

Why would you ever want to know about either of these relationships to practice archaeology or ancient history? While the production function concept helps us focus on the input-output

relationships of production in quantity terms, there may be times when the information we have relates to prices or costs (values, in general). We may want to have a clear grasp of what will happen to the production cost (the unit cost) of a community’s grain crop when the opportunity cost of one of their inputs (say, labor or land) appears, from the evidence at hand, to have changed. The profit function shows how we can move quite directly from the production-function concept to an output-cost relationship. Alternatively, we may have direct information on virtually no economic parameters but still have pressing research questions that involve essentially economic properties of artifacts. We can see quite directly that, say, Roman demand for agricultural slaves would have been depressed by falling agricultural prices. If we find evidence of sustained low prices, we have a good idea to look for pressure of one sort or another on slave holding.

2.11 The Economics of Mycenaean Vases, I: Supply and Cost

A recent study of Mycenaean vases at Ugarit, published together with a set of extended commentaries, has explored what the value of those vases might have been.¹⁰ The Dutch archaeologist Gert Van Wijngaarden approached the problem from the perspective of the contexts in which three categories of Mycenaean vases – small stirrup jars, amphoroid kraters, and conical rhyta – were found. Context is well suited to reveal information about the income of consumers of these items, but less well suited to reveal relative price information. In fact, Van Wijngaarden himself focused primarily on issues of consumer demand – personal valuation – which takes us to the following chapter rather than this one. However, this article incorporates such a wide range of economic issues regarding these vases that we will address parts of it in this and the next two chapters – those on demand and the structure of competition (market structure). Our treatment of the issue in this chapter will not deal directly with the issue that Van Wijngaarden and the three commentators addressed (with one exception) but rather takes a step backward from the question, “How much would the Ugartians have paid for these vases?” to the related

question, “How much would they have had to pay for them?” While this particular example is from the Late Bronze Age and, depending on your point of view, either the Near East or the Aegean, and about pottery, the principles the case brings up are far more general and could apply to any time or region, as well as to a much wider range of products. Fortunately, however, pottery is of primary importance to archaeologists, and many ancient historians have found their record quite valuable in their own endeavors as well.

The many interesting topics in demand that Van Wijngaarden (1999a, b) raised regarding these vases and their locally produced counterparts I will save for Chapter 3. In her commentary, however, Voutsaki (1999, 29) went beyond pure demand issues with the proposition that demand and supply of these vases could not be separated because the producers appear to have made their products for specific markets (29).¹¹

Let’s consider Voutsaki’s proposition, that Argive potters may have had Ugaritic (and Cypriot as well as others possibly) buyers in mind when they made their vases, thus linking (inseparably?) demand and supply. Look back to the form of a supply function and it’s obvious that both product prices are in it and factor costs are in it, either explicitly or implicitly. So the valuations of the Ugaritic and Cypriot consumers placed on these vases would indeed have affected the Argive potters’ supply of them. This does not mean, however, that those consumers’ full demand functions are so interwoven with the potters’ supply rules that the two concepts cannot be disentangled. In fact only the product price (specifically the f.o.b. price – “freight on board”; the price at the pottery gate, without shipping costs) is in the supply function, and as we will see subsequently, any specific price is jointly determined by both demand and supply. However, for any given schedule of factor prices facing the Argive potters directly, a higher valuation by the Ugaritic and Cypriot consumers (and any others as well, including local rubes with money) would raise the quantity of pots these potters would be willing to supply. Working in the other direction, however, for any given valuation on the parts of consumers, higher factor costs would cause the potters to reduce the quantity of vases they would offer at any given price.

On the other hand, the valuations of consumers could have had absolutely no effect on the potters’ production costs – look back to the cost function using just labor and capital earlier in this section. If the cost functions had the characteristic of constant costs – that is, the quantity produced does not affect the unit cost of production – then the only way the demand for these vases could have affected their production costs would have been by so increasing the demand for one or more of the factors used in vase production that they drove up the corresponding factor prices – skilled (or even unskilled) labor, clay, fuel. When we consider the likely relative magnitude of the LH III Argive potting industry relative to that economy’s agricultural sector (some readers may consider these anachronistic terms; they are a nomenclatural convenience, the many nuances of which will emerge gradually throughout the remainder of this text), it seems unlikely that unskilled labor used in potting would have had its wage driven up, but it is probably possible that skilled potters (or pot painters, if they differed) might have risen a bit or that high-quality clay deposits could have become more costly to mine. These are simply empirical questions with empirical answers.

Since we’ve already drawn some contrast between supply and cost, let’s take up an assessment of cost functions for these vases – what the customers in Ugarit might have *had* to pay for these Mycenaean vases, regardless of what they might have been willing to pay. Let’s work with a cost function for the vases in which the unit cost of each vase type is the sum of the input unit costs weighted by the cost shares for each factor: for example, the cost share of labor times the wage rate around Argos, plus the unit cost of clay times cost share of clay, plus the unit cost of fuel times the fuel cost share, and so on. Of course we don’t know any of the unit costs, but the cost shares might not be out of our reach, by virtue of contemporary ethnographic studies of potting. We might determine that the clay cost component of one of these vases probably increased just about in proportion to the volume of clay in the vase, making that component of a small stirrup jar’s cost much smaller than that of a large, amphoroid krater’s. However, we may have grounds for suspecting that the skill required in controlling a substantially larger volume of clay

would have been proportionally greater than that required for a smaller vessel of any shape; with such a suspicion, we would assign a higher labor cost share to the amphoroid kraters than to either the small stirrup jars or the conical rhyta. On the other hand, we might be led to believe that the skill required in manipulating the smaller, more restricted volumes and surfaces was greater than that required for a larger volume vase, which is more “forgiving” of errors. In this case, the labor cost share would fall with vase size. Similarly with firing costs: how do they increase in clay volume, as modified by vessel thickness and kiln size as dictated by vessel height? Although we may not know the answers to these cost-function questions, at least some of them probably are knowable, and we may begin to develop some elements of the functional forms of these vases’ cost functions. Comparing cost functions of vases made at different locations would be difficult because the relative labor, clay, and fuel costs may have differed substantially, but across vases made within the same region those costs shouldn’t have varied a whole lot, permitting reasonable comparison of costs of different types of vases.

2.12 Accounting for Apparent Cost Changes in Minoan Pottery

Changing the subject modestly to Minoan vases from the MM IIB through LM IA periods, Van de Moortel (2002) offers a detailed assessment of labor inputs at individual steps in the production of both fine-ware and utilitarian vases from various sites in central Crete. She reasonably equates the apparent input of labor time into a vase with labor cost. Under this definition, labor costs of utilitarian vases did not differ much from those of fine-ware in the Mesara, although in Knossian pottery, the expected lower labor input into utilitarian vessels was apparent. Labor inputs decreased substantially from MM IIB to MM III in all types of vases both at Knossos and in the Mesara. Labor time per vase in LM IA appears to have increased from MM III levels although not to MM IIB levels, but vases appeared far more standardized in size, proportions and decoration and in a smaller number of shapes. She quite

reasonably sees the standardization as the visible remains of a successful effort to cut costs.

Van de Moortel considers but rejects the possibility that the final destructions of the Protopalatial Period might have increased the demand for hastily made new vases, on the grounds that such an effect, if it did occur, should have been short lived, while the reduction of labor inputs into pottery lasted over three ceramic periods (192). However, she does not entertain the possibility that demands for construction labor, including cleaning up the Old Palaces’ rubble, during the initial phases of the Neopalatial Period, might have raised all wages (relative to the prices of other factors of production) over an extended period of time, prompting the observed reduction in labor inputs in pottery production in the face of an unchanged demand schedule for vases. She raises the possibility of increased competition among potters over this time period leading them to rush their work to cut costs, but also suggests a change in the demand for quality in all vases, at a time of no evidence of decreased prosperity. The only explanation offered for an increase in competition in pottery production is a reduction in or elimination of “administrative regulation” of pottery after the Protopalatial Period (206). The introduction of new shapes should be indicative of changes in demand for variety (see Chapter 3, section 11), possibly associated with changing social rituals, but attributing the quality deteriorations that occurred to changes in consumers’ expectations (206) – essentially, tastes – is less convincing than regarding the introduction of the clumsily shaped conical cups as representing a genuine change in demand caused by changing social practices. Suggesting “administrative deregulation” (206) as sharing responsibility for either increased competition or simply shoddier work conflicts with the evidence she cites regarding the distribution of Kamares ware throughout the Knossos area (204–205).

That lower labor inputs would produce vases of lower total cost need not be the case however. First, if wages increased, a reduction in labor inputs might leave the cost of a typical vase unchanged. Other major production steps that still involved labor, but other inputs as well, are acquisition and preparation of the clay and firing,

and she notes evidence of less preparation time in the MM III pottery (195). The unit cost (wages) of labor time in vase construction (call that L_3) need not have been the same as the unit cost of labor time in clay acquisition (L_1) and preparation (L_2) if different individuals with different skills were involved; my own guess about relative wages (w_i), assuming different individuals undertook these tasks, which might not have been the case, would be $w_3 > w_2 > w_1$, although an experienced potter might have a different assessment. Labor was certainly used in firing (L_4), but was probably small relative to the other labor inputs; whether the individual supplying L_4 was the same person supplying L_3 is an open question. I can see no reason for major nonlabor cost changes in clay acquisition or preparation, which seems to leave the fuel involved in firing as the only other nonlabor input whose cost might have changed over these periods, possibly because of deforestation. I have not cited all the nuances of Van de Moortel's findings, including differences in the timing of appearance of various changes between Knossos and the Mesara, but the simplest explanation I see (or hypothesize) for these changes is a lengthy period of increased demand for labor caused by both the destruction of the Old Palaces (clean-up) and the lengthy construction of the New Palaces, with some institutional learning involved to standardize many aspects of production and save clay expenses. Altogether, however, Van de Moortel's analysis makes some major strides in the economic analysis of Minoan

pottery production, which could provide a model for other economic analyses of archaeological remains.

2.13 Production in an Entire Economy: The Production Possibilities Frontier

So far we've considered production from the perspective of the individual producer, whether that agent is a single individual or a firm consisting of many individuals. We've noted that the expansion of supply by a single producer might be hindered by the efforts of other suppliers in the same industry to expand at the same time. The various producers might bid up the prices of resources that are particularly important in their industry. They might bid against producers in other industries who want to use some of the same resources. At any rate, the industry supply curves might rise more sharply than the individual firms' supply curves. The production possibilities curve is a diagrammatic tool from general equilibrium analysis that lets us consider directly how the expansion of production of one good cuts into the ability to produce another good, considered on an economy-wide (country-wide) basis. How do efforts to expand manufacturing production curtail the ability to produce agricultural products?

Figures 2.15(a) and (b) show isoquants for two goods, A and B, with fixed factor supplies

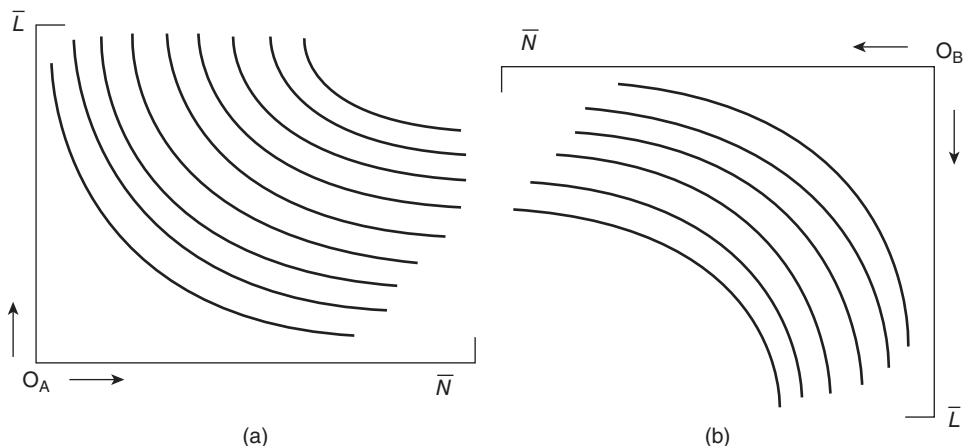


Figure 2.15 (a) Isoquants of good A. (b) Isoquants of good B.

available – $\bar{N}$ and $\bar{L}$. You could think of New Kingdom Egypt as the economy or country, agricultural output and royal construction as goods A and B, and labor and land as N and L . The analysis highlights the overall tradeoff between agricultural production and royal building.¹² In Figure 2.15(a), we start with the origin of the diagram conventionally in the lower left corner to characterize the production of good A, but we draw the isoquants for good B with the origin in the upper right corner. Both diagrams have the lengths of their axes defined by the fixed quantity of labor and land in the economy. Assume that goods A and B are the only two outputs in the economy, that land and labor are the only two inputs, and that we are looking for combinations of A and B production that will fully employ both factors. Figure 2.16 shows the two isoquant diagrams overlaid, in a diagram known as an Edgeworth–Bowley box. The curved line from O_A to O_B (called the contract curve) connects all the tangencies between A and B isoquants. Along this line, the marginal rates of substitution between land and labor are the same in both “industries,” which we know is a condition for efficiency in production. The areas between the overlapping isoquants contain feasible combinations of goods A and B, but not efficient ones (see the shading between the second good-A isoquant and the third-highest good-B isoquant).

The production levels associated with each isoquant in Figure 2.16 can be transferred to another diagram with the quantities of goods A and B on the axes, as in Figure 2.17. The curve

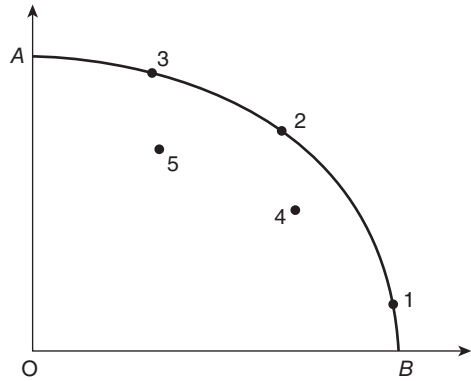


Figure 2.17 An economy’s production possibility frontier (PPF).

in Figure 2.17 is the production possibilities frontier. Points on the curve are points on the contract curve. All points inside the production possibility frontier (PPF) are feasible production combinations, although only those on the frontier are efficient. Points outside the PPF cannot be attained with the technology available to the economy under study. Points 1, 2, and 3 on the PPF in Figure 2.17 correspond with points 1, 2, and 3 along the contract curve in Figure 2.16. Points 4 and 5 in Figure 2.17 are inside the PPF; their locations in Figure 2.16 show their locations off the contract curve.

Figure 2.18 extracts one bit of information contained in the Edgeworth–Bowley box diagram – the factor input ratios for the two industries corresponding to point 2 on the contract curve.

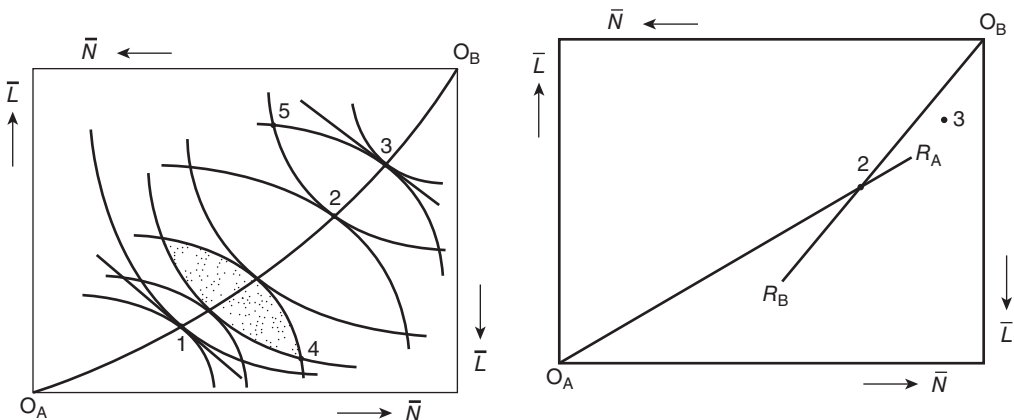


Figure 2.16 The Edgeworth–Bowley box diagram.

Figure 2.18 Expansion paths of outputs in an economy with fixed quantities of inputs.

At that combination of A and B output, industry A uses the higher land / labor ratio. Point 3 from the contract curve is identified, although, in the interest of keeping the diagram a little cleaner, lines representing the input ratios for that output combination are not drawn. It can be seen that the land / labor ratio of industry A would be even higher than it is at point 2, and so would that of industry B: as we move toward origin O_B , which would represent specialization in good A, the factor input ratio in the expanding industry approaches the factor availability ratio of the entire economy (an intuitively obvious result). Looking back at Figure 2.16, we also can see that the equilibrium ratio of factor prices changes as we move along the contract curve. Factor price lines are drawn at points 1 and 3, showing a higher relative price of land at point 1 than at point 3. Good A is the labor-intensive product (higher ratio of labor to land in production); when its demands on the fixed labor supplies are less intense (its production is quite low at point 1), the relative price of labor is lower than it would be if the output of good A were at point 3. Expansion of the output of the labor-intensive industry (which gets less labor intensive as it expands) drives up the relative price of labor.

Turning back to the production possibilities frontier in Figure 2.17, we can ask where the economy would choose to produce and why. The information on the diagram contains no information presently that would make one efficient point stand out over any other. Figure 2.19

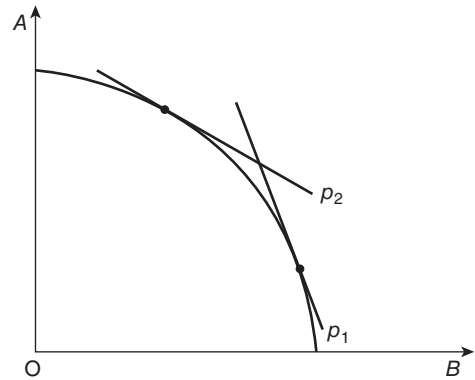


Figure 2.19 Alternative product prices on the production possibility frontier.

introduces relative price lines to the PPF diagram to give a reason to produce one or another combination of A and B. Relative price p_1 represents a high relative price of good B, and we see that the economy decides to produce a large proportion of B, relative to its capacity for producing B, at that relative price.¹³ At relative price p_2 , which indicates a higher relative price of good A than does p_1 , the economy produces more A and less B. We have not addressed yet why either relative price should appear. For that we must turn to the analysis of consumption and demand. The Edgeworth–Bowley diagrammatic apparatus contains much more information than we have extracted from it yet, but we should wait for the development of other analytical tools before taking them out.

References

- Cassidy, John. 1998. "The Political Scene. Monicomics 101." *The New Yorker*, September 21, 73–77.
- Conison, Alexander. 2012. *The Organization of Rome's Wine Trade*. Ph.D. dissertation. University of Michigan, Ann Arbor MI.
- De Mita, Jr., Francis A. 1999. "The Burden of Being Mycenaean." *Archaeological Dialogues* 46: 24–27.
- Duncan-Jones, Richard. 1982. *The Economy of the Roman Empire: Quantitative Studies*, 2nd edn. Cambridge: Cambridge University Press.
- Potts, D.T. 1997. *Mesopotamian Civilization; The Material Foundations*. Ithaca NY: Cornell University Press.
- Powell, Marvin A. 1999. "Wir müssen unsere Nische nutzen: Monies, Motives, and Methods in Babylonian Economics." In *Trade and Finance in Ancient Mesopotamia* (MOS Studies 1), Uitgaven van het Nederlands Historisch-Archaeologisch Instituut te Istanbul LXXXIV, edited by J.G. Dercksen. Leiden: Nederlands Instituut voor het Nabije Oosten, pp. 5–23.
- van de Mierop, Marc. 1987. *Crafts in the Early Isin Period: A Study of the Isin Craft Archive from the Reigns of Išbi-Erra and Šu-Ilīšu*. Leuven: Departement Oriëntalistiek.
- Van de Moortel, Aleydis. 2002. "Pottery as a Barometer of Economic Change: From the Protopalatial to the Neopalatial Society in Central Crete." In *Labyrinth Revisited; Rethinking "Minoan" Archaeology*, edited by Yannis Hamilakis. Oxford: Oxbow Books, pp. 189–211.
- Van Wijngaarden, Gert-Jan. 1999a. "An Archaeological Approach to the Concept of Value: Mycenaean

- Pottery at Ugarit (Syria)." *Archaeological Dialogues* 46: 1–23.
- Van Wijngaarden, Gert-Jan. 1999b. "The Value of an Archaeological Approach: A Reply." *Archaeological Dialogues* 46: 35–39.
- Voutsaki, Sofia. 1999. "Value beyond Ugarit." *Archaeological Dialogues* 46: 27–31.
- Whitelaw, Todd. 1999. "Value, Meaning and Context in the Interpretation of Mycenaean Ceramics." *Archaeological Dialogues* 46: 31–35.

Suggested Readings

- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapters 7–8.
- Varian, Hal R. 1992. *Microeconomic Analysis*, 3rd edn. New York: Norton. Chapters 2–5.

Notes

- 1 On one of the alternative paradigms used to study ancient behavior – that of Karl Polanyi – the recent retrospective assessment by Powell (1999) is interesting.
- 2 For a careful assessment of cost and price data from the Roman Imperial period, see Duncan-Jones (1982, Chapters 2 and 3).
- 3 Notice that although we can pick the size of the change in N (or generally, in our "control variables"), we have to take whatever the consequent change in Q (or more generally, the "objective") is, because that relationship is given to us by our technology.
- 4 I have not used a Cobb–Douglas production function for this demonstration because, with constant returns to scale, its supply function is not well defined.
- 5 "Own-price" is a technical term used to distinguish between the effects on an input or an output of changes in its own price and changes in the prices of other goods or inputs, which can still have an influence on the quantity demanded of the good or factor in question. When considering the effect of the price of some good or factor a , the term "cross-price" is used.
- 6 Which, contrary to some apparent beliefs expressed in the ancient historical and archaeological literature, is simply some branches of neoclassical economics in which components of models, which in other instances would be fixed as assumptions, are modeled as endogenous – as Conison (40) recognizes.
- 7 Expressed alternatively, the MRTS must equal the ratio of marginal products.
- 8 Note that we're not saying that these marginal products *will* be equalized across products. However, because the producer (owner of the firm) will make more of whatever it is she gets paid in when she combines her productive inputs efficiently, there will be forces within the firm to accomplish that. When those internal forces are insufficient by themselves, forces external to the firm (competition for inputs in the variable input case – and competition in the supply of identical or substitutable products) will reinforce or replace the internal forces. In other words, there are reasons to believe that the firm will be somewhere close to its productive "optimum" in its factor allocation.
- 9 You'll also notice that the level of output, Q , doesn't show up on the right-hand side of this expression, which means that this producer can produce as much of this output with this technology as she wants without driving up her cost. This is an artifact of the constant returns to scale built into the production function by virtue of the output coefficients, α and $1-\alpha$, summing to 1. If they summed to less than 1, unit cost would rise with higher levels of output; if they summed to more than 1, unit cost would fall as output increased. Another way of getting an increasing unit cost as output rises is for the unit costs of the inputs to rise as more units of them are used in production: this producer would have to bid them away from other uses, and as the output levels of those other products fell, the prices consumers would be willing to pay for them would rise, raising the marginal products of the inputs in those uses, and accordingly the amounts those other producers would be willing to pay for them.
- 10 See also the exchange with De Mita (1999) and Whitelaw (1999).
- 11 She also raises the perpetual question about higher quality Bronze Age ceramics from the Aegean, "Was it produced 'under central palatial control?'" While some aspects of this issue are fruitfully studied as topics in market structure (the structure of competition), we will find it useful to return to it again in Chapter 6, on public economics, under the subject of the economics of planning.
- 12 Some readers may suggest that Pharaoh would have built during the agricultural slack seasons.

This may have been true to some extent, but the tradeoff still occurs because building would therefore be restricted. To loosen such restrictions on construction, he would have had to accept restrictions on agricultural production.

- 13 Again, some readers may wonder what kind of a “price” royal construction could have had. This is a complicated issue in public goods – the subject of Chapter 6. Suffice it to say here that some implicit

kind of a price would have existed for the public (royal) construction in terms of agricultural output, regardless of whether the royal house thought in those terms or not. If this result bends the mind too much, you can recast the entire analysis as an example of, say, agriculture and privately produced and owned manufactured goods, items that could have been priced easily and explicitly.

Consumption

While supply provides one blade of the famous Marshallian scissors, demand forms the other, and as two blades of a pair of scissors are required to cut paper, supply and demand together are required to determine values. One of the common, first-line objections among scholars of the social sciences and the humanities to using economic models to assist their thinking is that those models impose artificial assumptions of rationality on the people it studies, and people just aren't rational. While this attitude may be *prima facie* evidence for the contention itself, it may repay us to lay out first (section 3.1) the exact assumptions about the consistency of behavior embedded in contemporary economic models. Economists have, however, considered the charge seriously and have gone some way toward introducing models of choice by agents who act irrationally (section 3.14), but one of the principal lessons from this research has been that, if we want to claim irrationality on the part of our subjects, to make intellectual progress we must specify (and model) the type of irrationality we believe to exist. The other major finding is that the results of choices motivated by irrational behavior look an awful lot like the results of choices motivated by rational behavior, so getting excessively

particular about rationality may not make all that much difference.

The principal goal of this chapter is to show students of the ancient Mediterranean world enough about how the contemporary theory of consumption behavior – demand theory – is constructed and operates (how it works and why) that they can apply the concepts flexibly to problems encountered in their daily studies of the ancient world. Accordingly, the first seven sections of the chapter are relatively bite-sized building blocks, which, individually, tend to be rather abstract, not lending themselves easily to convincing examples from the ancient world: the concept of the finite budget that places limits on behavior (section 3.2); measuring wellbeing (section 3.3); the demand function, which is a paragon of a technical lexicon drawn misleadingly from everyday language (section 3.4); subsequent properties of the demand function – that is, the restrictions on choices that it implies (section 3.5); getting from a single individual's demand for goods to the total demands of lots of people who compete with one another for fewer material goods than they all would like (section 3.6); and using the demand concepts to identify, and possibly measure, changes in wellbeing (section 3.7). We will draw on the principles from these sections,

plus those from other sections of the chapter, in section 3.17 to continue our examination of the economics of Mycenaean vases, which we began in Chapter 2.

In between sections 3.7 and 3.17, however, the topics become more intuitively correspondent to problems with which we are familiar in our own daily lives, and we can find ancient parallels more readily. For example, section 3.8 on price and consumption indexes assembles the principles of the preceding sections to address the issue of comparing prices or consumption bundles across locations, times, or both, asking whether people at one time or place seem to be better off in a clearly definable sense. For instance, with all the price data that have been assembled from, say the Athenian Agora or Ptolemaic Egypt, what do we have to do to interpret them sensibly?

Intertemporal choice (section 3.9) was a ubiquitous problem facing ancient farmers: how much grain to retain as seed for the subsequent planting and how much to eat? Certainly rules of thumb were developed, and farmers did not make complicated calculations each season, but both the ancient farming literature and contemporary scholarly research on ancient farming leave more the impression of rigid formulas than rules of thumb subject to choices around the edges in response to the notorious fluctuations in agricultural success. While food is consumed once, many goods can be consumed over much longer periods of time (section 3.10): clothing, tools, houses. Nonetheless, consumers must pay for many of these durable goods when they acquire them (payment on time having been more restricted than it is today until quite recently), and being more long lasting, many of these items cost more than simple consumer goods. How do people work these larger allocations into their everyday lives, and with what consequences? The issues in durable goods lead naturally into the matter of product variety (section 3.11). The simple, expositional treatments of demand implicitly or explicitly deal with homogenous goods, but even many goods that look the same to noncognoscenti have considerable variation within the categories that “outsiders” can blur:

varieties and qualities of cereal grains, qualities of wines and oils, cloth and clothing, jewelry, and on and on. Section 3.11 introduces several ways of organizing thinking about these characteristics of products.

People’s time links virtually all of their activities, from working in fields to eating meals to caring for children, to making love and war. Simple possession of agricultural products does not feed one. Converting grain in storage pithoi into a meal on a table requires the use of time as well as other material inputs such as cooking implements and fuel for heat. The household production model has been the workhorse economic vehicle for studying allocation decisions outside explicit markets and connecting them to events within markets (section 3.12). Many decisions are made with full knowledge that the decision maker only stands a certain chance of getting the hoped-for outcome, as contrasted to the feared outcome. Uncertainty is virtually ubiquitous, and people adjust their behavior to try to reduce it, mitigate its impacts, or shift its burdens to their fellows. Section 3.13 sketches outlines of the basic issues involved in behavior under uncertainty. We’ve already introduced the explicit modeling of irrational behavior, which section 3.14 treats.

The final three sections of the chapter apply these concepts to issues of the ancient world. Section 3.15 explicitly addresses the possibility of fixed and immutable prices (or at least sluggish ones) such as are claimed frequently for the ancient world. It considers what would be required to keep agricultural prices in particular as stable as some claims would have them and questions the factuality of the observations, reminding the reader of the difference between absolute and relative prices. Section 3.16 takes advantage of some demand relationships to make a rough estimate of the possible range of family incomes at Pompeii in the first century C.E. Finally, section 3.17 continues investigation of the economics of Mycenaean pottery begun in Chapter 2 from the cost perspective. In this chapter’s installment we apply the basic concepts of the first half of the chapter, then use some

of the more specialized perspectives such as durable goods and product variety. The purpose of that section is not to answer quantitatively or in some other definitive manner the question of the value of Mycenaean vases in the late second millennium Near East, but to offer a simple and consistent framework for posing the questions archaeologists have been asking for a while now. With luck, it will point scholars in those fields to extant (if not necessarily easily accessible) data of which they are aware, as well as to strategies for collecting data that may not be collected routinely in current excavation and cataloguing strategies.

3.1 Rationality of the Consumer

The use of rationality as an assumption about behavior by economics has been the subject of considerable rock-throwing among scholars in other social sciences and the humanities. In fact, the economist's definition of rationality is quite restricted, consisting of only three parts. First, if a consumer is confronted with any pair of alternatives, she either prefers one of them to the other or is indifferent between them. This is known as the Completeness Axiom. The second part of rationality is that if the consumer prefers alternative *A* to alternative *B*, and alternative *B* to alternative *C*, then she also prefers alternative *A* to alternative *C*. This is known as the transitivity axiom, and it simply says that tastes are consistent – not necessarily admirable, but consistent.¹

Can tastes change over time? Of course. Would we be willing to explain consumption behavior, either of the same person over time or between individuals at the same time on the basis of changed (or different) tastes? If we were to allow that, it would be an admission of our total ignorance and ability to predict. The third part is that if the consumer chooses some alternative *A* from a group of alternatives which we call Ω , then she prefers *A* to all the other items in Ω other than *A*. This is called the rationality axiom, and it is the basis of what is called “revealed preference.” Note that this axiom does not allow for the possibility

that our consumer “really prefers” *B* but chooses *A*. Preference is based on observed behavior rather than declared interests.²

3.2 The Budget

A difference between our treatments of production and consumption is that we have no hesitation at all in assigning consumers budgets. Several things about consumer behavior can be learned from examination of the budget itself. The budget (sometimes called the budget constraint) represents the fundamental condition of scarcity facing the consumer. It is what she has available to spend. We can represent a budget in two dimensions, representing the income available to a consumer who can spend it on two goods, *A* and *B*. Figure 3.1 shows good *A* on the ordinate and good *B* on the abscissa, with a line joining the two axes representing all the maximum combinations of the two goods that our consumer can acquire with the income defined by this budget line. The negative of the slope of the budget line, $(-p_B/p_A)$ represents the relative prices of goods *A* and *B*. We do not need money to denominate these goods' prices. Consider the following expression of how the consumer can spend her budget (income): $p_A A + p_B B = Y$. If we divide

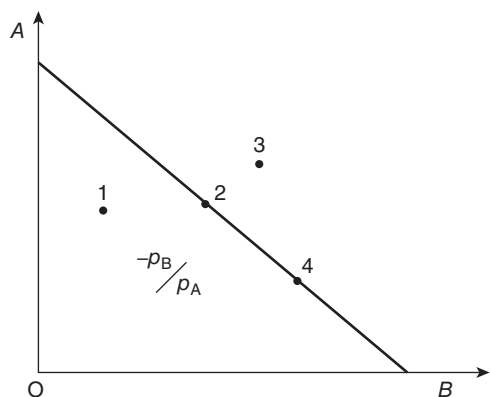


Figure 3.1 A consumer's budget and possible choices.

both prices and money income Y by the price of good A , we have $A + (p_B/p_A)B = Y/p_A$; we have used the price of good A as the numeraire, and money income also is denominated in terms of good A . Working with such relative prices is referred to as studying the economy in “real” terms: prices are actually denominated in terms of one of the physical goods under study. Y/p_A is called “real” income, but it is only real income in terms of good A . An alternative definition of real income would use good B as the numeraire.

The consumer in Figure 3.1 could choose the combination of A and B represented by point 1, but in doing so would not expend the income fully. Since we have not included savings, this leaves some expenditure unaccounted for, and the consumer can do better for herself. Points 2 and 4 are both reachable with this budget, but point 3 is not. Figure 3.2 shows the budget line with the original relative prices and the same budget with changed relative prices. (How do we know that the money income is the same? We can still purchase the same amount of good A if we were willing to spend all our income on one good.) This change in relative price could represent an exogenous change in the supply of A or the imposition of a tax on good A . A shift in the budget line from ab to $a'b''$ represents an increase in the budget.

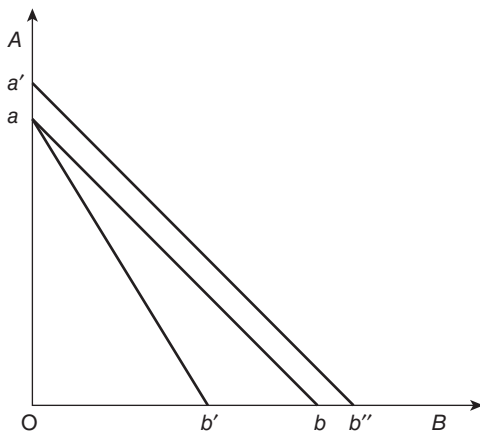


Figure 3.2 Alternative relative product prices and budgets facing a consumer.

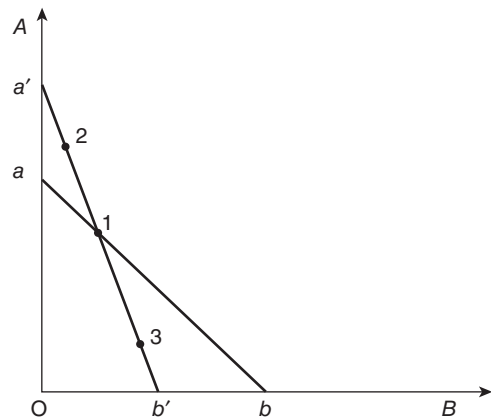


Figure 3.3 Consumption choices under alternative budgets with different relative product prices.

In Figure 3.3, the consumer makes the initial choice of combination 1 at the budget described by budget line ab . Suppose the relative prices and budget change to $a'b'$. Will the consumer choose point 2 or point 3 in response to this new income and set of relative prices? We have enough information to be able to answer this question. At the original set of relative prices, our consumer could have chosen point 3 but chose point 1 instead. At a new set of relative prices which permit her to choose either 3 or 1, since she has already demonstrated her preference for 1, she will not switch to 3 now. At the original income and set of prices, point 2 was unattainable, so her choice of point 1 does not conflict with her choice of point 2 under a changed set of prices. Since the relative price of good A has fallen, the consumer will under most circumstances move toward point 2 on the new budget line, at which she will consume a higher proportion of the good whose price has fallen, but under no circumstances will choose a consumption point which she has already revealed as less preferred to a choice still available.

3.3 Utility and Indifference Curves

Utility as economics has come to use the concept is purely notional. In the nineteenth century

and into the early twentieth, many economists considered utility a cardinal concept – that is, quantities that would bear continuous numerical comparison, such as “15 is three times as large as 5.” In contemporary economic theory the concept of utility is a description of consumer tastes and yields an ordinal measure of wellbeing: level 2 of utility is greater than level 1 and lower than level 3.

Once again in a two-good context, a utility function of an individual consumer is specified as $U = f(A, B)$, where U is the ordinal level of utility, f represents the function, and A and B are the quantities of goods consumed during the time period in question. In this function, a change in the consumption of, say, good A changes the level of utility; increases in consumption raise utility and decreases reduce it. The amount by which a small change in the consumption of good A , the consumption of all other goods remaining the same, changes utility is called the marginal utility of A . We can express this algebraically as $\Delta U / \Delta A$ (or give it the acronym MU_A) for any good which is really a “good” (that is, it is desirable rather than undesirable), marginal utility is positive (U changes in the same direction as A). An important property of the utility function is that, just as in the production function, changes in utility from subsequent increases in the consumption of any good, holding constant the quantities consumed of all other goods, are negative (we could write this as $\Delta(\Delta U) / \Delta A^2$, which is read as “the change in the change in U , divided by the change in A multiplied by itself”). This reflects decreasing marginal utility: the more of something we have, the less we value further increments of it. Another important property of the utility function is what is called the “cross-effect” between pairs of goods: the change in the marginal utility of good A with a larger consumption of good B ($\Delta(\Delta U) / \Delta A \Delta B$). These cross-effects are generally positive, but they could be zero. The case of zero cross-effects in consumption (technically the case in which the ratio of marginal utilities of two goods are unaffected by changes in the consumption of a third good; this condition is called “separability”: the first two goods are separable from the third in

this particular case) has important implications in a number of uses of utility theory; we will see one such case in the construction of price indexes below.

The most common graphical exposition of the utility function uses indifference curves, which are iso-satisfaction lines comparable to isoquants in production theory. Figure 3.3 shows a family of indifference curves that could be associated with the utility function just presented. The set of (A, B) pairs on indifference curve I_0 all yield the same level of satisfaction. The level of satisfaction derived from indifference curve I_1 is greater than that from I_0 but less than what could be derived at level I_2 . Indifference curves are parallel. If they were not, at some point they would cross, which would violate the second axiom of rationality. The slope of the indifference curve at any point identifies the marginal rate of substitution between goods A and B , just as the slope of the isoquant told us the marginal rate of technical substitution between inputs.

We can overlay the budget line on the indifference graph and show the relationship between income and the level of indifference (satisfaction) achievable. Figure 3.5 shows indifference curves I_0 through I_3 , all parallel to one another. Suppose our consumer starts with income indicated by budget line Y_0 . She can do better than reach the level of satisfaction associated with I_0 . I_0 cuts the budget line in two places: at point 1, the marginal rate of substitution of A for B exceeds the relative

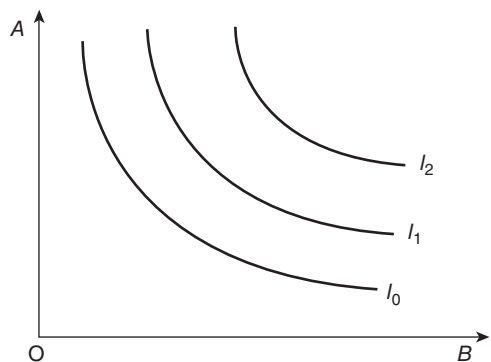


Figure 3.4 Indifference curves.

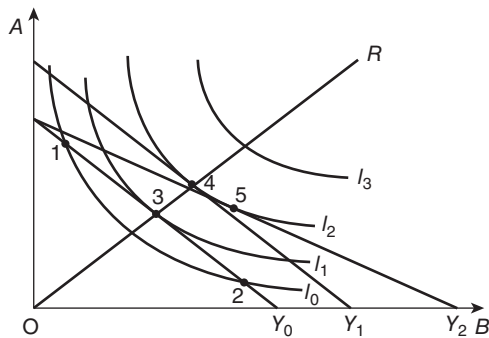


Figure 3.5 Budget lines and indifference curves determine consumption choices.

price of A and B , and point 2, it falls short of the relative price. In light of the rate at which she can acquire the two goods, the consumer will find a higher level of satisfaction with less A and more B than at point 1 and more A and less B than at point 2, although both points are feasible with her income. By making these substitutions, she raises her level of satisfaction to I_1 at point 3. Suppose her real income rises to Y_1 . The relative price of A and B remain constant, so she chooses the same ratio of the two goods at point 4, where she enjoys satisfaction level I_2 . Now, while she is on I_2 , the relative price of A and B changes as reflected in the new budget line Y_2 . (How do we know that the real income associated with money income Y_2 is the same as the real income embodied in Y_1 ?) At the new relative price, she consumes more of good B , which is now cheaper relative to good A , and less of good A .

Both A and B are “goods” in the strict sense of the term – they are desirable consumption items such as food and clothing. Not everything that consumers deal with is desirable, however. The ones that aren’t are called, quite intuitively, “bads.” Some examples: garbage or other household refuse and, especially interesting, risk. In the case of a good and a bad, indifference curves slope upwards, as in Figure 3.6. The relative price line also will slope upwards, although the analogy to a fixed budget cannot be derived from the upward sloping relative price line. Preferred levels of indifference lie upward and to the left, as I_1 lies above I_0 . We might think of good A as food

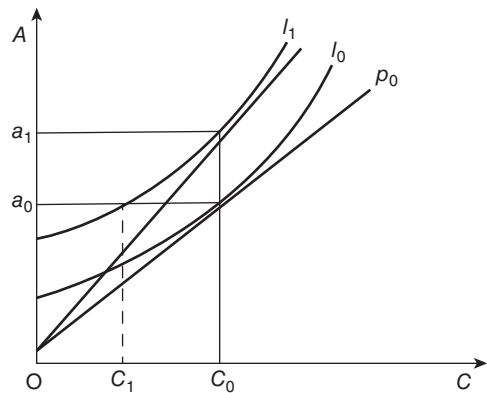


Figure 3.6 Indifference curves and consumption when one “good” is a “bad.”

and “bad” C as a measure of the consumption risk associated with food. At the relative price of food and risk (the cost of getting rid of some additional risk, associated with the acquisition cost of food) represented by p_0 , the consumer accepts c_0 amount of risk (we will define the units of risk in a later chapter) and a_0 food. Consuming the same quantity of food would let the consumer reach indifference level I_1 if she were able to divest herself of $c_0 - c_1$ risk. Alternatively, if risk remained at c_0 , it would take $a_1 - a_0$ additional food to let her reach satisfaction level I_1 . Another way of looking at these compensations, if our consumer had $a_1 - a_0$ amount of food taken away from her, she would feel no worse off if $c_0 - c_1$ risk were also removed.

The Cobb–Douglas and CES functional forms are handy to characterize utility functions. Empirical work on consumption is conducted with demand functions, or related concepts such as the expenditure or cost function. We will consider functional forms for those constructs but we will not need them for further exposition of the utility function.

3.4 Demand

The concept of demand relates the quantity of a good purchased in a given time period to (i) the price of the product relative to the prices of other products, (ii) the prices of substitute

and complementary products, and (iii) income. As was the case with production, there are several ways to develop the theory, each consistent with the others. First, we can derive a demand curve graphically from an indifference diagram. Figure 3.7 shows indifference curve I_0 , along which are drawn three relative price lines, p_0 , p_1 , and p_2 . Starting at relative price p_0 , the consumer with the indifference system represented by I_0 will consume a_0 quantity of good A (don't worry how much of good B he consumes; we're working on the demand curve for good A). Price line p_1 raises the relative price of good B, and our consumer buys (consumes) relatively more of good A – quantity a_1 . Now, let's experiment with a reduction in the relative price of good B with price line p_2 . The consumer buys quantity a_2 of good A since its relative price went up so sharply. Now, take this information on quantities purchased at various relative prices to Figure 3.8, where we will build the demand curve implied by the information in Figure 3.7. The axes on the demand diagram are the relative price of good A on the ordinate and the quantity of good A (Q_A) purchased at that price. First, notice that the price on the ordinate is the relative price of good A – in terms of good B, because that's the only other product we're dealing with right now. At relative price p_2 , we plot quantity a_2 ; at relative price p_0 , we mark off quantity a_0 ; and at price p_1 , we plot quantity a_1 . These three points in p - Q_A space form a downward sloping line if we connect them, which we do. We can label this demand curve D_0 , since it is associated with the level of

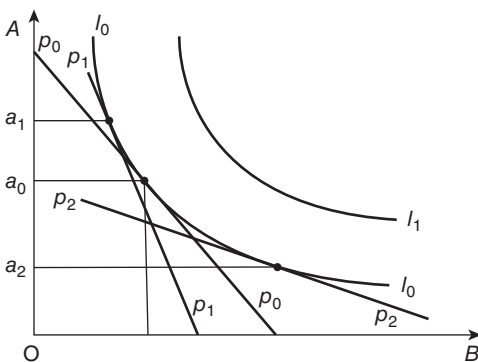


Figure 3.7 Indifference curves and budget lines as income changes.

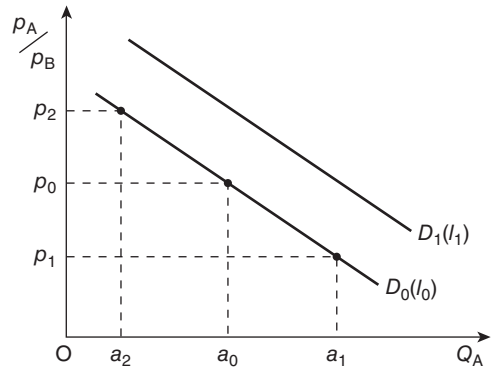


Figure 3.8 Demand curves emerging from budget and indifference schedules.

real income associated with indifference curve I_0 . If we were to draw another indifference curve on Figure 3.7 that represented a higher real income level (curve I_1 on Figure 3.7), we can see, without drawing all the relative price lines (actually expanding each of the lines p_0 through p_2 parallel to their counterparts at the indifference level I_0) that at the identical relative prices, the consumer will purchase more of good A (as well as more of good B). These consumption points underlie demand curve D_1 in Figure 3.8. We have just seen that an increase in income will shift the demand curve out to the right. Reading this information alternatively, we see that a consumer with a larger income would be willing to pay a higher price (relative price) for any given quantity of a good than he would have been willing to pay out of a smaller income.

It will be useful to show several alternative ways to derive a demand function. The procedure we have used in Figures 3.7 and 3.8 has the consumer minimize the expenditures required to reach a given level of utility. Formally, this problem is: Minimize $E = p_A A + p_B B$, subject to $U^* = U(A, B)$. Using a Cobb–Douglas utility function ($U = Q_A Q_B$) for the general form $U(A, B)$, and deriving the first-order conditions with respect to goods A and B and the Lagrange multiplier on the utility constraint, we can solve for the demand functions for both goods A and B as $Q_A = q_A(U^*, p_B/p_A)$ and $Q_B = q_B(U^*, p_A/p_B)$. Both demand functions have the properties that (i) as the relative price of the good in question

risers, the quantity demanded falls (this property of a demand curve is known as the fundamental theorem of demand theory) and (ii) as the level of utility achievable rises, the quantity demanded rises. We can write the general form of this demand function as $Q_i = f(p_i, p_j, U)$, where the subscript i represents the good whose demand is described (making p_i the “own-price”) and j indicates the relative price of any other products which might affect the demand for product i . As p_i goes up, Q_i falls; as p_j goes up, Q_i can either rise or fall, depending on whether good j is a substitute or complement (but if there is only one other good in the demand function, it has to be a substitute; with multiple goods j , they cannot all be complements). This demand curve is known as the “compensated demand function,” or the “Hicksian” demand curve. What it is compensating for is the possibility that relative price changes can affect the real income that any money income can yield; by maintaining the level of utility constant, it compensates for these real-income changes created by price changes.

Next, consider what is actually an older demand function, the “ordinary” or “Marshallian,” demand function. It is derived from a different optimization problem imposed on our consumer – maximizing utility with a given income. This problem is: Maximize $U = U(A, B)$, subject to $Y^* = p_A A + p_B B$. The consumer spends his entire income (abstracting from savings) on goods A and B in whatever combination will yield her the highest level of utility. Using the Cobb–Douglas form again, we take the first-order conditions for maximization of this problem by manipulating the quantities of goods A and B and the Lagrange multiplier on the budget constraint, and solve for Q_A and Q_B . We get the following ordinary demand functions for goods A and B : $Q_A = Y^*/2p_A$ and $Q_B = Y^*/2p_B$. This ordinary demand function demonstrates an important property of demand curves in general, known as “homogeneity of degree zero in all prices and income.” If we multiplied both income and prices by the same number, the multiple would cancel, leaving demand unchanged.

A useful feature of the ordinary demand function is that it can be converted into an inverse demand function, in which price is a function of quantities and income rather than quantity being a function of prices and income. This function is derived from the utility maximization problem, but we simply solve for price instead of quantity. In this formulation, the price of good A is the ratio of its marginal utility to the sum of the marginal utilities of all other goods; in the two-good model we have been using, this boils down to $p_A/p_B = MU_A/MU_B$, which is simply the tangency condition between an isoquant and a budget line that we’ve been using. Now we see how it comes from a maximization problem of the consumer. Inverse demand functions are particularly useful when prices either do not exist or are distorted.³ In such situations, the prices they yield are called “shadow prices,” and may not be the same as any prices that are observed for the goods under study. These shadow prices will be a more reliable guide to the resource costs of goods than are observed (or “market”) prices when the latter are subject to considerable government interference or are generated by poorly functioning private markets. A common example of “poorly functioning” private markets comes from contemporary developing countries (and in many industrialized countries not so long ago): The “market” for rural income insurance is virtually nonexistent. Farmers could try storing storable food grains and hope they get to them before the rats and bugs do; they could try to “obligate” some near or distant neighbors or relatives to feed them during an especially poor crop year if the neighbors and relatives don’t experience the same crop conditions, and so forth. Another alternative among tenant farmers (renters, either for shares or fixed rents) is to combine their rentals of farm land with implicit obligations to be helped out by the landlord during a bad year. In such cases the rental on land will be higher than the marginal product of land, or in the case of share rentals, the landlord’s share will be on the high side. In either case, the rental price of land will not reflect either its marginal cost to its supplier (the landlord)

or its marginal benefit to the renter because it is conducting transactions in two separate “goods” – land and insurance. In such instances, as well as many others in which it has been difficult to establish markets for particular goods and services, a tool that will estimate shadow prices of the goods under study will help reveal the real resource costs that decision makers are accepting.

Another manipulation of the ordinary demand function yields what is called the “indirect utility function.” The quantities of the consumption goods that maximized utility in the derivation of the ordinary demand curve are functions of prices and incomes. We can substitute the prices and incomes that maximized utility into the utility function for the quantities of goods in that function. This gives a different version of the same utility function, one in which utility is a function of prices and income rather than the quantities of goods consumed. In doing this, we have tucked the income constraint, which was external to the original utility function, into the indirect utility function. Sometimes it will be more convenient to use an indirect utility function than a direct one because prices and incomes are usually more readily available than are consumed quantities of goods. The indirect utility function has the form $U = g(p, Y)$. Higher values of prices depress utility, and higher values of income raise it.

A final manipulation of these demand relationships yields the expenditure function (sometimes called the cost function, but not to be confused with the producer’s cost function from production theory – although they are structurally quite similar). We can arrive at it in either of two ways. First, we can invert the indirect utility function, which comes from the ordinary demand function: E (expenditure, which equals income, Y) = $g^{-1}(p, U)$ (read this as “ g inverse of p and U ”). Alternatively, we can substitute the functional structure of prices and utility that minimized expenditures to give the compensated demand function, into a cost function, in a fashion comparable to the way we substitute prices and income for quantities of goods to get the indirect utility

function. This gives us $Q = h(p, U)$, in which Q falls when prices rise and increases for higher levels of utility.

3.5 Demand Elasticities

The budget in consumer theory imposes a number of necessary relationships on the responses of consumption of various goods to changes in prices and income. These relationships offer considerable predictive power about consumer choices. These relationships are expressed quite well in the form of various elasticities. Recall from the discussion of production theory, that an elasticity is the percentage change in a dependent variable divided by a 1% change in an independent variable. Consider a general form of an ordinary demand function: $Q_i = f(p_i, p_j, Y)$. The “own-price” elasticity of demand for good i is $(\Delta Q_i/Q_i) \div (\Delta p_i/p_i)$, frequently designated $\varepsilon_{ii} < 0$ (it is always negative). The income elasticity of demand for good i is $(\Delta Q_i/Q_i) \div (\Delta Y/Y)$, frequently designated $\eta_i > 0$ (positive for all “normal” goods; can be negative for “inferior” goods like various types of foods that poor people eat because they can’t afford anything more palatable; think twice before assigning a negative income elasticity to a good when you’re modeling some problem). The “cross-elasticity” of demand for good i with respect to the price of any good j is $(\Delta Q_i/Q_i) \div (\Delta p_j/p_j)$, frequently designated ε_{ij} . It can be positive (for substitutes), negative (for complements) or zero (for completely unrelated goods). When there are several goods that we identify as j (for example, when we say “ $j = 1, \dots, n, j \neq i$ ”), they all may be substitutes for good i , but they may not all be complements; at least one must be a substitute, and more than likely, quite a few will be substitutes.

The first relationship is that the weighted sum of the income elasticities of demand must equal 1: $\sum s_i \eta_i = 1$, where s_i is the share of expenditures on good i in all expenditures ($p_i Q_i/Y$). The second relationship places constraints on the values that own- and cross-price elasticities of demand for any good can take in an ordinary

demand function: $s_i \varepsilon_{i1} + \sum_j s_j \varepsilon_{ij} = -s_i$. This is called the Cournot aggregation condition. For compensated demand functions, this condition is $s_i \xi_{ii} + \sum_j s_j \xi_{ij} = 0$. The difference between the two relationships arises because the compensations implicit in the $-s_i$ term on the right-hand side of the Cournot condition have already been included in the compensated demand elasticities ξ_{ii} and ξ_{ij} .

Probably the most famous relationship in demand is the Slutsky equation, which decomposes an own-price elasticity of demand into substitution and income components. This relationship is $\varepsilon_{ii} = \xi_{ii} - s_i \eta_i$. For cross-price elasticities, the Slutsky equation is $\varepsilon_{ij} = \xi_{ij} - s_j \eta_i$. Figure 3.9 shows the Slutsky decomposition graphically. Indifference curves I_0 through I_2 represent the utility function (the Slutsky equation comes from the problem $\text{Max } U = U(A, B) \text{ s.t. } Y = p_A A + p_B B$) underlying the demand function for which the Slutsky terms are generated. The initial relative price of goods A and B is represented by price line ab . Now that relative price falls to what is represented by price line ac , which represents a cheapening of good B. The consumer originally chose the combination of goods A and B at point 1 to maximize her utility. At the new relative prices, and the real income they imply, she chooses point 3 on indifference curve I_3 . This full change of consumption choices contains a pure substitution effect and a pure income effect. If we draw a price

line parallel to the new price line, but through the original consumption point (line dd), the consumer would want to change her consumption by substituting some more of good B for some of what she presently spends on good A, at point 2. This effect alone would let her gain a higher level of satisfaction, represented by indifference curve I_1 . The change in consumption attributable to the effective increase in income created by the decrease in the price of good B is represented by the move to the higher, parallel budget (price) line ac , which accounts for the full effects of the price change. The consumer's movement from consumption point 2 to consumption point 3 is the income effect of the price change. The move from point 1 to point 2 is represented by the substitution term in the Slutsky equation, and the movement from point 2 to point 3 by the weighted income term.

Income elasticities of demand yield particularly interesting information. All normal goods have positive income elasticities (η), but those with $\eta > 1$ will experience an increase in their budget shares as income rises, while the budget shares of those with $\eta < 1$ will fall as income rises. If $\eta = 1.0$ for some good, its budget share will remain constant with changes in income. The value of the income elasticity of demand can be used to divide goods into the categories of luxuries, necessities, and "inferior" goods. Goods with $\eta > 1$ can be called luxuries, those with $\eta < 1$ (but still positive) necessities, and those with $\eta < 0$ inferior goods. Recall, however, that the budget-share-weighted sum of income elasticities of demand for all goods equals 1. Consequently, goods with particularly large budget shares are unlikely to have either especially high or especially low income elasticities.

Another concept related to the income elasticity of demand, yet distinct from it, is the Engel curve, which is the relationship between consumption of a particular commodity or commodity group and income. Engel curves can be estimated (they are primarily empirical relationships, although the logic of demand theory places some restrictions on them) for individuals or for entire populations ("the market"). The Engel relationship can take two principal functional forms: $Q_i = f(Y)$ or $s_i = f(Y)$. The former relates the quantity

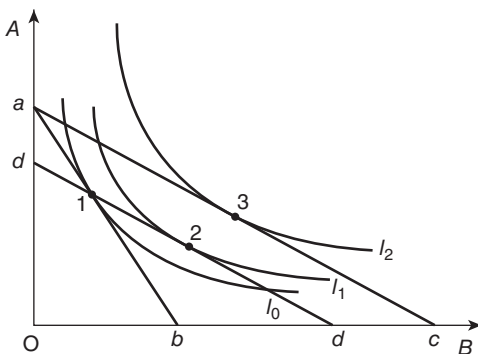


Figure 3.9 Income and substitution effects on consumption of product price changes.

purchased of a commodity to the income of the consuming group or individual, whereas the latter relates the share of expenditures on a commodity to income. Either relationship can be estimated for individual households or entire populations or subpopulations. In the Engel relationship, the sign of the coefficient on Y , for example, the term b in the equation $Q = aY^b$, $b > 0$ indicates luxuries and $b < 0$ necessities. The distinction between necessities and inferior goods is a matter of magnitude and hence some element of arbitrariness.

A final fact about demand elasticities, both price and income: long-run elasticities are generally larger in absolute value than short-run ones. Parallel to the discussion of cost and supply responses in Chapter 2, section 2, “long” and “short” sound like they refer to lengths of time, which is a reasonable association, but technically they refer to the ability of consumers to adjust all their consumption choices: the long run is the situation in which all of those adjustments have been made. The short-run is the period of the initial impact, when only those that can be made immediately have been made. There are, of course, many “intermediate” runs involving different degrees of adjustment. Short-run elasticities typically are estimated (econometrically) from data on different individuals’ (or other observations’) responses at pretty much the same time. Long-run elasticities typically are estimated from observations of the same consumption units over a number of time periods. As can be imagined, assembling the time series sufficiently lengthy for long-run estimates adds complications that I won’t go into here.

Having pulled you through these conceptual thoughts, you may well be wondering what order of magnitude typical (“typical”) demand elasticities possess. Estimation procedures certainly can affect the magnitude of estimates (all demand elasticities are estimates), and in recent years, econometric methods for estimating parameters from time-series observations have made considerable strides, and time series are one of the two major sources of demand elasticity information. Choices of estimation methods can make statistically significant differences in

estimates, but frequently not economically significant, although it is not impossible that a method may push its estimate across the economically important value of 1.0. Even then, the difference from 1.0 might not be considered economically significant. With this preface, I can offer two references on magnitudes of demand elasticities for a number of products and services, both of which are old at this point – but this kind of study does not appear to be conducted currently. First, Houthakker and Taylor (1970, Chapter 3) reports own-price and income (expenditure) elasticities for a number of goods and services using post-World War II United States data and their Chapter 5 (217–227) reports elasticities from a different estimation method for the United States, Sweden and Canada. Second, Lluch *et al.* (1977, 53–64) report demand elasticities estimated for a variety of products for a number of developing countries with data from the mid-1950s through the late 1960s. In both sources, of course, some of the contemporary products cited did not exist in antiquity, although some of these might have had at least rough parallels. At any rate, these two sources can give an idea of the order of magnitude of a number of income, own-price, and sometimes cross-price elasticities of demand for specific products or product groups.

3.6 Aggregate Demand

The theory of demand as we have developed it so far pertains to the individual demanding unit – individuals or households. Frequently we want to know how the demand for, say, wheat, across the entire economy would change as various prices and income changed. We *can* add individual demand curves horizontally, just as we did with individual supply curves: $D_i = \sum_n f(p_i, p_j, Y)$, in which D is aggregate demand for good i and we sum over n individuals. However, because of differences in individual tastes and differences in incomes across individual demanders, we cannot construct an aggregate demand function of the form $D_i = f(p_i, p_j, \sum_n Y)$ and have the intricate restrictive properties of the

individual demand function carry over. It is certainly the case, however, that many an estimation of aggregate demand curves for specific products has been estimated empirically, using functional forms along the lines of $D_i = ap_i^{e_{ii}} p_j^{e_{ji}} M^{\eta_i}$.

Despite the limitations on aggregation of individual demand curves, we now have derived the apparatus with which to study aggregate demand for individual commodities. Some technically inexact simplifications prove extremely useful in the predictions they provide. For instance, although we know that not all individuals have the same tastes, and that they observably have different incomes, we are almost always safe in relating a change in population (the value of n in the aggregate demand equations of the previous paragraph) to an increase in aggregate demand, if not necessarily by a factor of n . Similarly, if we know that total income in the economy has increased, possibly as much because of increased income per capita as from population increase, we generally are safe in believing that the aggregate demand curve will shift up and to the right, if not necessarily by a factor equivalent to the percentage increase in aggregate income. When we are searching for *qualitative* results as contrasted with quantitative (“did some variable increase or decrease?” rather than “by how much?”), knowledge of this order can be quite useful.⁴

3.7 Evaluating Changes in Wellbeing

A problem commonly faced in the economic evaluation of changes in an economy is deciding whether the population is better or worse off for the change. Changes in price, technologies, and government policies frequently present us with situations whose welfare effects we want to evaluate. Ordinal utility theory and the compensated demand curve offer useful analytical vehicles to make such evaluations. From utility theory, the twin concepts of the equivalent variation and the compensating variation offer a basis for developing a monetary measure of changes in wellbeing (or equally one denominated in some numeraire good). We will show the equivalent

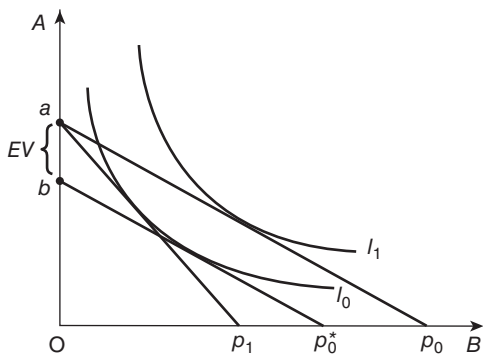


Figure 3.10 Evaluating changes in wellbeing 1: equivalent variation.

and compensating variations diagrammatically, then show how they are translated to the demand curve, and finally discuss some practical issues in measuring the relevant quantities.

In Figure 3.10 the price of good B rises from p_0 to p_1 as a consequence of having a tax imposed on it. With the income represented by budget line p_0 , the consumer was initially at indifference level I_1 . With the price change, her effective income is described by budget line p_1 , and she reaches the lower indifference level I_0 . At the old prices but the new level of indifference, her income in terms of good A is represented by the intercept of budget / price line p_0^* , at point b on the ordinate. The quantity of good A represented by the distance ab on the ordinate represents the income in terms of good A that the consumer would be willing to pay to avoid the tax. This amount of income in terms of either money or a numeraire good (good A in this case) is called the equivalent variation. To find the equivalent variation, we draw the old price line tangent to the new indifference level. The equivalent variation is known also as the “willingness to pay” either to avoid a “bad” or to acquire a new “good.”

The compensating variation identifies how much income our consumer would have to be given to make her as well off after the tax as before. In Figure 3.11, we start again with price p_0 and consumption that lets our consumer reach indifference level I_1 . The tax increases the price of good A to p_1 , which puts our consumer on indifference curve I_0 . To keep her at the same

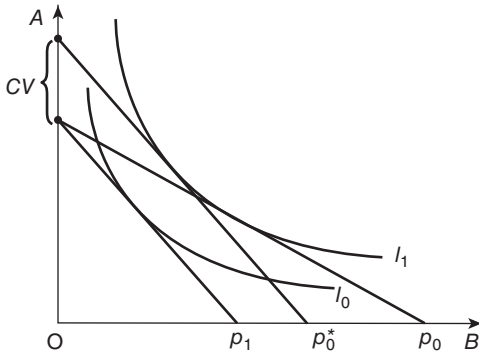


Figure 3.11 Evaluating changes in wellbeing 2: compensating variation.

indifference level with the new, tax-inclusive price, she would have to be given income in the amount of CV on the ordinate. To find the compensating variation, we draw the new price line tangent to the old indifference level. The combination of substitution and income effects used to identify the equivalent and compensating variations should be familiar from our previous examination of those same effects in the analysis of utility and demand.

It is possible to translate these results to the compensated (Hicksian) demand curve – the one that keeps utility constant rather than income. Using the indifference curve, it is possible to identify the quantity of good A that the consumer would be willing to forego to obtain the quantity of good B associated with any relative price of goods A and B . Start at price p_0 in Figure 3.12. The consumer is willing to offer a_0 units of good A for the very first unit of good B (identified as b_0). For the next unit of good B , b_1 , the consumer is willing to offer a_1 units of good A . And so on until she reaches the prevailing price represented by price line p^* , at which price she will offer a^* units of good A for the b^{*th} unit of good B . Each of these units of good A that the consumer would be willing to offer for units of good B less than quantity b^* , are called marginal valuations of the a_i^{th} unit of good B . The shape of the indifference curve guarantees that marginal valuations of successive units of good B in terms of good A will be smaller than the marginal valuations of previous units. These decreasing units of B that would

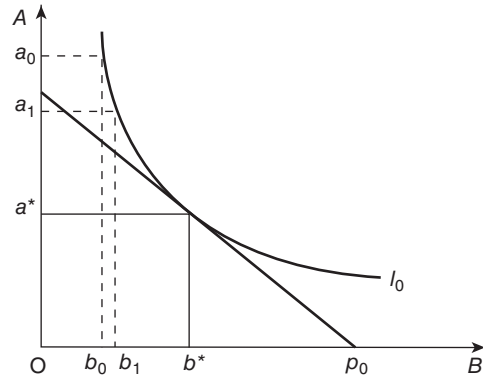


Figure 3.12 Compensating variation with a constant-utility demand curve.

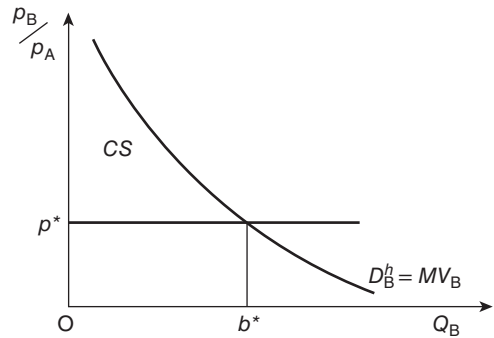


Figure 3.13 Consumer's surplus.

be given up for an extra unit of good A display the decreasing marginal valuation of good A (as more is consumed). This appears in Figure 3.13 as a declining compensated demand curve, or marginal valuation curve, for good B . We label the demand curve D_B^h to identify it as the Hicksian compensated demand curve, and also identify it as MV_B , the marginal valuation curve for good B . The area under the demand curve but above the price p^* , is known as consumer's surplus.

To follow the consequences of a price change in terms of consumer surplus, Figure 3.15 shows a price increase from p^* to p^{**} . The area $p^{**}a_1a_2p^*$, which has the dimensions of price times quantity, or revenue, and could be measured either as money or in terms of a numeraire good, is the compensating variation of the price change.

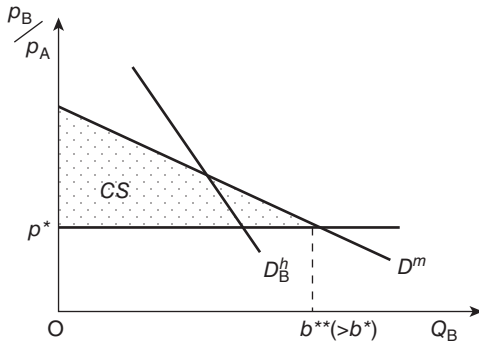


Figure 3.14 Changes in consumer surplus following a product price change.

Compensated demand curves are difficult to observe empirically because they depend on unobservable utility, and in practice we are generally left observing and measuring ordinary (Marshallian) demand curves, which depend on income. We show the ordinary demand curve, D^m , and a compensated demand curve based on the same utility function, in Figure 3.14; consumer surplus is shown as the shaded area. The difference this makes in our welfare measures appears in Figures 3.30 and 3.31. If we repeated the construction of marginal valuations of good A in terms of good B in Figure 3.12 for higher indifference curves, we would derive higher marginal

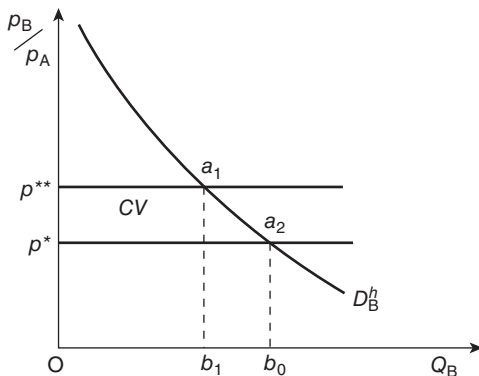


Figure 3.15 Consumer surplus with ordinary (Marshallian) and compensated (Hicksian) demand curves.

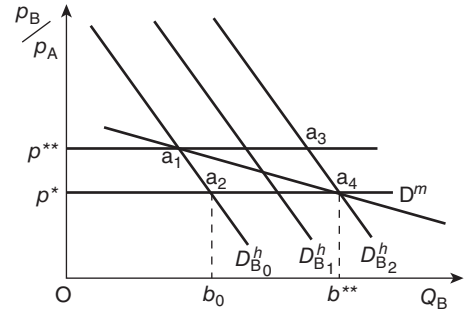


Figure 3.16 Marshallian (ordinary) and Hicksian (compensated) demand curves.

valuation curves, representing marginal valuations of successive units of good A at different levels of utility. Without repeating the graphical derivation, we show the series of MV curves as a series of Hicksian compensated demand curves in Figure 3.16. The ordinary demand curve does not hold utility constant, and the same level of money (or numeraire-denominated) income will include several levels of utility, represented by the multiple Hicksian demand curves. Thus, the empirically observed demand curve will have greater slope (have greater price elasticity; be more sensitive to price) than will the compensated demand curve. In Figure 3.16, area $p^{**}a_3a_4p^*$ is the equivalent variation, while the area $p^{**}a_1a_4p^*$ is consumer surplus. Thus, we have the relative magnitudes $CV > \Delta CS > EV$.

Empirically, the differences between these measures are not especially large, possibly in the order of 5%. We will unabashedly prefer to work with the consumer surplus measure simply because it is more accessible empirically and consequently easier.

While we are on the subject of “surpluses,” the concept of a “producer’s surplus” should be addressed. It is common to refer to the area above the supply curve of a product (or industry) and below the price line as producer’s surplus, like the crosshatched area apb in Figure 3.17. In some circumstances this area will represent a surplus comparable to consumer’s surplus above the price line but below the ordinary demand curve, but not always, and possibly not even frequently. If S_B is the supply curve of a competitive manufacturing industry, area apb will not be a surplus;

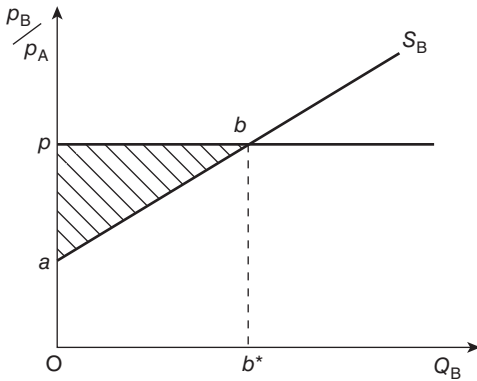


Figure 3.17 Producer's surplus.

any rents earned by factors such as labor and land will be included in the area of resource costs under the supply curve between the origin and b^* . In fact, expansion of output by the industry producing good B could cause the *erosion* of rents (as contrasted to opportunity costs; refer to the discussion in the chapter on production) to factors of production it uses itself as well as increases in rents to factors used elsewhere; if such expansion caused rents to increase to some factors used in industry B , rents to that factor would increase in any other activities using their services elsewhere in the economy. A perfectly discriminating monopsonist would be able to capture area apb as a rent, but such market power is rare.⁵

Consider an application of consumer surplus analysis to the problem of improving transportation costs between producers and consumers of some good. Let the problem be one of producers of, say, grain, spatially separated from the consumers of at least the part of the grain not consumed on-farm. The example could refer to the supply of grain from the countryside to cities as well as to long-distance transportation of virtually any commodity. Figure 3.18 shows the demand curve for grain as perceived by the consumers; it is labeled D_{cif} ("c.i.f." being an abbreviation for "cost, insurance, freight," and meaning delivered price).⁶ The supply curve that the producers see at the farm gate is S_{fob} ("f.o.b." representing "freight on board" but not delivered – that is, produced but still waiting to

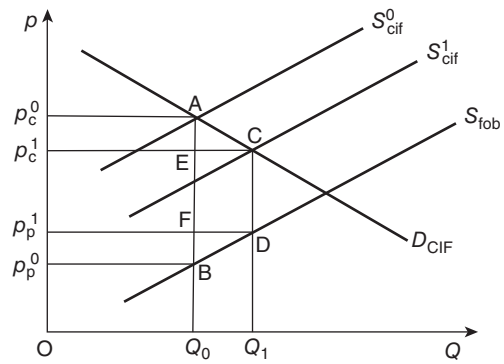


Figure 3.18 Transportation costs 1: costs perceived by producers and consumers.

be distributed; otherwise known as "farmgate prices"). The producer has nothing to do with transportation (we assume this; at any rate, if the producer and transporter are the same person, the activities are different). The supply curve of grain delivered at the consumers' location is S^0_{cif} ; it contains shipping costs. The transportation cost of a unit of grain is the vertical distance AB . Producers supply quantity Q_0 at the farmgate (f.o.b.) price they can get (p_p^0), and consumers are willing to purchase that amount at price p_c^0 . Now, suppose that the transportation rate falls to an amount per unit of grain equal to distance CD . The demand (think of it as the "demand schedule") for grain in the city is unchanged; it remains at D_{cif} because no "shifter" variables in the demand function, such as income or other prices, have changed. Similarly with the supply curve of the producers; none of their technological conditions have changed, none of the input costs they face, and so on, so they are operating on the same supply curve, S_{fob} . But, the delivered supply curve has fallen to S^1_{cif} . Consumers in the city are willing to purchase the larger quantity Q_1 at the lower delivered price, p_c^1 , and the farmers receive the farmgate (f.o.b.) price of p_p^1 for this quantity. Consumer surplus for the city consumers rises by the amount $p_c^0ACp_c^1$; they have to pay a lower price for the original quantity Q_0 , equal to area $p_c^0AEP_c^1$ and would be willing to pay amount AQ_0Q_1C for the additional grain, $Q_1 - Q_0$, but only have to pay price p_c^1 per unit for it, leaving them with additional consumer surplus equal to area AEC .

As we have just noted before we embarked on this example, it frequently is difficult to associate the area under the price line (p_p^1 in this case) and above the supply curve with a producer surplus comparable in interpretation to consumer surplus, but consider in this case that farmers face a flat supply curve for labor and all additional price per unit of output is rent (in the sense of a surplus, not a land price per period). Then area $p_p^1 DB p_p^0$ is producer's surplus – or at least additional income to producers. Of this increase in income, farmers receive the additional amount $p_p^1 FB p_p^0$ on the amount they originally supplied and area FBD on the incremental supply. Both producers and consumers are better off from the decrease in transportation costs (improvement in transportation technology).

We can take an alternative look at this problem. Figure 3.19 looks at the difference between demand as noticed by farmers in the countryside and consumers in the city and relies on an f.o.b. supply curve – that is, the farmgate supply curve as seen by the farmers. In this view, farmers perceive the demand for their grain to have shifted to the right, to D_{fob}^1 , by virtue of the improvement in transportation technology. With this shift in demand, they are willing to supply quantity Q_1 at unit price p_p^1 , with unit transportation cost of CD (down from amount AB). With an unchanged demand for the delivered grain (D_{cif}), urban consumers see the combination of the increased supply and reduced transportation

cost as delivered price p_c^1 . Although farmers are getting paid more for this additional supply, they are paying a lower unit price altogether for both what they consumed prior to the transportation improvement and their incremental consumption following it.

3.8 Price and Consumption Indexes

We frequently want to compare how well off a consumer or group of consumers is at different points in time or at different locations. A consumer may consume different quantities of goods and face different relative prices at different times or locations, so the comparison takes on some complexity. Similarly, we may want to track the progress of prices over time, individually or in groups. If all prices changed at the same rate, this task would be simple, but generally, relative prices change, and if the composition of goods consumed also changes, we may have some uncertainty about our prices measures. Price indexes and cost-of-living indexes let us perform these comparisons in systematic ways that we can tie back to basic consumer theory. The three basic types of index are price indexes, cost-of-living indexes, and real consumption indexes.

There is no unique, “best” price index, although some are superior to others because of their reliability in capturing what we want them to capture. The two simplest price indexes follow the development of the compensating and equivalent variation measures of welfare change just discussed. In creating an index of prices of different commodities, we need to find a weight for the commodity prices. The natural weight is the quantity of each good, and the resulting index is the ratio of the weighted prices in some later period to some base period. However, we have the choice of comparing the weighted prices using the consumption pattern of the base period or the later period. The Laspeyres index uses the early, or base, period consumption. Suppose we have just two periods, 0 and 1, and just two goods, a and b . The Laspeyres index of the price level in period 1 relative to period 0 is $P_L = (p_a^1 q_a^0 + p_b^1 q_b^0) \div (p_a^0 q_a^0 + p_b^0 q_b^0)$. Figure 3.20 shows what the Laspeyres index is measuring.

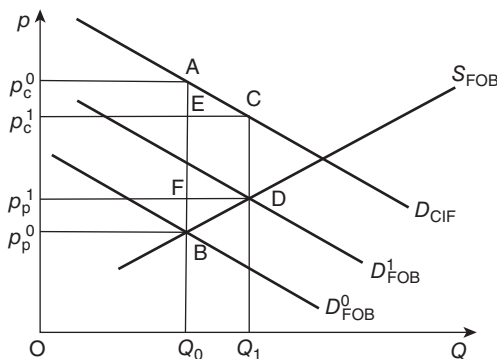


Figure 3.19 Transportation costs 2: demand as viewed by farmers in the countryside and consumers in a city.

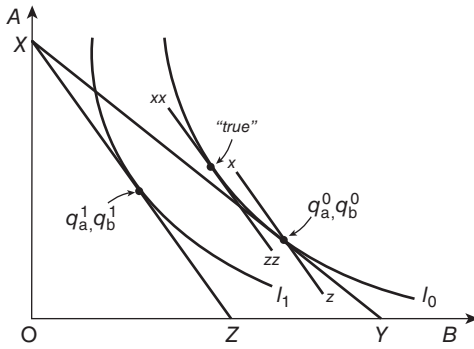


Figure 3.20 The Laspeyres price index.

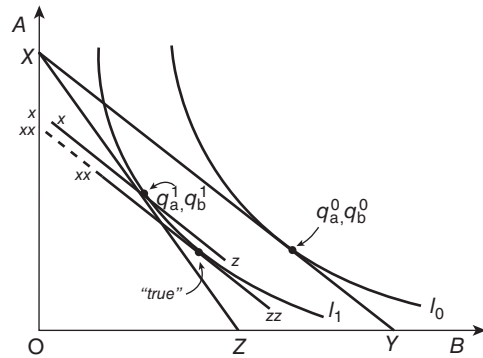


Figure 3.21 The Paasche price index.

The original set of prices is represented by budget line XY , to which indifference curve I_0 is tangent at the consumption combination q_a^0, q_b^0 . In period 1, the price of good B increases, rotating the budget line clockwise to XZ , where indifference level I_1 is reached at the consumption combination q_a^1, q_b^1 . We draw the new relative prices through the original consumption point with line xz . Note, however, that the new price line is drawn *through* the original consumption point, so the Laspeyres index does not keep utility constant. The “true” index, that is, one that kept utility constant would be represented by the tangency of this same, new price line, $xx-zz$, to I_0 , up and to the left along the indifference curve. Thus, at the new set of prices, the consumer could reach the same level of wellbeing as before more cheaply by altering his combination of consumption goods. The Laspeyres index thus has a slightly greater value than the true index of prices.

The other common index number, the Paasche index, parallels the compensating variation, and its construction is shown in Figure 3.21, which reproduces the same period-0 and period-1 prices and the same indifference system. The Paasche index uses the later period’s consumption as the weights in a price index, represented by budget line xz , which in this case is the original relative price line transferred to the new indifference level. Its formula is $P_p = (p_a^1 q_a^1 + p_b^1 q_b^1) \div (p_a^0 q_a^1 + p_b^0 q_b^1)$. In this case, the true index – that is, the one that would keep utility constant – would use a consumption combination down and to the right (toward the good that is cheaper in the new

relative price regime). The Paasche measure is distance Xx on the vertical axis, which is smaller than the distance associated with the true measure, $X-xx$.

All the data necessary to construct both the Laspeyres and the Paasche indexes can be observed readily. The true index, based on the utility function, requires estimation of the consumption combination at the constant-utility tangency, and with a large number of commodities this quickly gets expensive in practice. Consequently, much effort has been devoted to deriving formulas that use only readily observed data (prices and consumption quantities) that approximate the true index. One of the most popular is the Törnqvist index, which is a weighted-average, logarithmic index: $\log P_T = \sum_i \frac{1}{2} (w_i^0 + w_i^1) \log(p_i^1/p_i^0)$, in which the w_i^t are the budget shares of goods i in period t , and p_i^t are the commodity prices in period t . It is not a constant-utility index. In practice, estimation with contemporary data indicates that the Laspeyres and Paasche indexes frequently yield values within a half percentage point of the estimated true index. If there were no substitutability in consumption, the Laspeyres, Paasche, and true indexes would all be identical. In a practical sense, the Laspeyres index is somewhat easier to use than the Paasche because the latter requires continual updating of the base consumption weights if one wants to extend the index beyond one period later than the base period.

What does one *do* with a price index once one has constructed one? Use as a deflator is quite

important in comparing prices of individual goods at different times. Direct comparison of the price of, say, a unit of wheat at two different times, without reference to the general level of prices at each time period, as reflected in a price index of one period in terms of the other, conveys uncertain information. If the level of prices had been absolutely stable over the intervening time, the comparison would indeed tell what happened to the “real” price of wheat. If the level of prices rose or fell over the time interval, the simple, undeflated comparison would be meaningless; we would not know if the real price of wheat rose or fell, regardless of what the comparison of undeflated prices says, and since knowing the answer to that question would have been the motivation of the inquiry in the first place, we are left with absolutely no information on the original question despite having some data. Undeflated comparisons of individual prices between or among locations at a single time period may be safer to make than comparisons over time, but only inasmuch as tendencies for real prices to be equalized to within transport costs between locations are realized.⁷ Of course, if one suspects that the cost of living (COL) differed between some of the pairs of locations being compared, direct (undeflated) price comparisons yield merged information on both the relative prices of the good in question and the local costs of living, and we have no way of disentangling the two sets of information other than site-specific COL deflation.

A cost-of-living (COL) index measures the relative cost of reaching a given level of utility (standard of living) in different circumstances, either temporal or spatial. In addressing the temporal question, the COL index fixes the consumption pattern and measures the cost of purchasing that particular commodity basket in different time periods. Thus, the consumption pattern still forms the weights for the prices. Using the Laspeyres and Paasche approaches, the consumption pattern of comparison could be either the base or the later (“current”) period. A common comparison of interest is the urban-rural or city I-city J cost-of-living comparison. For this interlocational COL index, the consumption basket can be that of either location, and the prices

reflect the cost of consuming that basket in the different locations. A slightly more complicated comparison would trace the costs of living in alternative locations over time, in which case the base consumption basket is both a locational and temporal base.

A real consumption index compares two different living standards, across either time or space. In a consumption index, in contrast to a price or cost index, the prices are the weights attached to the different consumption bundles. If we were to compare real consumption in, say, eighteenth-century B.C.E. Egypt and Mesopotamia – simplifying the consumption bundles to some major portion of the consumer budget for which data could be obtained, say, food and housing, or possibly food, clothing and housing – we would use consumption data from Egypt and Mesopotamia and weight those bundles by the prices of one of the locations.⁸ With the Egyptian prices, the index would tell us the cost of consuming the Mesopotamian consumption bundle at Egyptian prices relative to the cost of consuming the Egyptian bundle at Egyptian prices. The opposite comparison would simply switch the prices from Egyptian to Mesopotamian. It might well prove that consumers in one region would prefer the other region’s relative prices of some of the major commodities.

Subindexes for either the cost of living or real consumption contain some subtleties that should not be ignored. A subindex would consider only a portion of the array of consumption goods, say food or housing – subsets of consumption that frequently are of especial interest themselves. It is easy to select only a partial array of consumption goods and calculate the ratio of later-period to base-period prices / costs or quantities, but referring back to the basics of utility theory we see that the quantity chosen of any particular good depends not only on its own price, but those of other goods. If we segmented, say, food consumption from the rest of consumption and created a cost index for obtaining a fixed combination of food items and watched the value of that index change over time, the actual combination of food items selected over time could well be affected by changes in prices outside the scope of

our subindex, and a Laspeyres or Paasche index calculation would diverge even further from the true measure of cost change than would be the case with the full array of consumption goods. Only if the goods in the subindex are separable (in the sense discussed above under the subject of utility theory) from those omitted, will this interdependence be avoided. Additionally, subindexes cannot be weighted and added together to yield a total index because of the same problems with interactions between included and excluded commodities.

Price changes – and price differences among locations – can affect different groups of consumers differently. Over some time periods, price changes might hurt poor consumers and benefit wealthy ones; vice versa at other times. Family composition affects consumption, so price changes and differences can have differential impacts across categories defined by family composition as well as across income groups.

Calculation of these indexes over long periods can yield problematic results. Some goods may drop out of production, new ones may enter, and the qualities and characteristics of others may change. What is a “long period”? That, fortunately or unfortunately, is a judgment call of the student.

3.9 Intertemporal Choice

So far, we have considered a consumer's decisions within the time horizon of the present. Let's consider a present period now, but one in which the consumer is aware of a future period as well. Intertemporal choice necessarily introduces the interest rate, for which a number of non-mutually-exclusive theories exist. We will use two of these theories about interest rates to show different aspects of the intertemporal choice problem. The first involves productive activities, and the second involves what has been called a natural rate of time preference in consumption.

Under many circumstances, the consumer will have a choice of how much to consume this period and how much to consume next period, with additional consumption this period

decreasing the amount available to be consumed next period. In one such circumstance, our consumer might produce something that grows over time. If he starts with a stock S_0 of it this period, and consumed absolutely none of it, by next period it would have grown to the amount $S_0(1 + g)$, where g is the growth rate of the stock. Suppose he took a look at his stock and decided to simply divide the present amount in half, so he could consume half ($S_0/2$) this period and half next period ($S_0/2$). His plans would go awry, because next period, he would have the amount $(S_0/2)(1 + g)$ instead of ($S_0/2$). Would he be disappointed or satisfied? Would he have consumed more last period if he had realized how much he was going to have in the subsequent period?

We've already reached ahead of ourselves, so let's slow down and consider our consumer's preferences for consumption this period and next period. Instead of considering different commodities in a utility function, let's consider a utility function that contains consumption in general at different dates: $U = f(c_0, c_1, \dots, c_n)$, where the subscripts refer to time periods, the present period being period zero. To simplify the analysis, we will consider only two periods – this period and next. Figure 3.22 shows a set of indifference curves between present and future consumption. Recall the interpretation of the slope of each point on these indifference

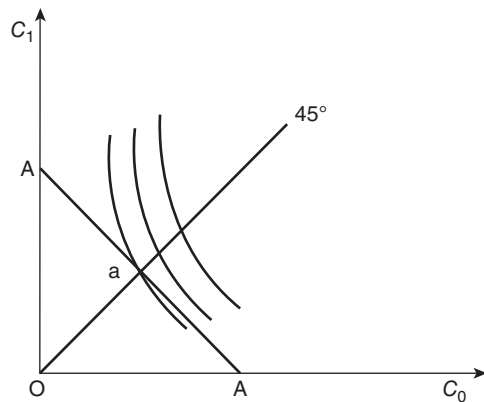


Figure 3.22 Choice between present and future consumption.

curves: it is the rate at which the consumer would be willing to trade a unit of one for a unit of the other. We have the shapes of these indifference curves drawn in a particular configuration: “budget line” AA has a slope of -1 , representing equal “prices” of present and future consumption, or a relative price of 1 of present and future consumption. We have drawn a ray from the origin at 45° , which cuts the unit-price budget line at point a. One of the indifference curves cuts through point a, without being tangent to the unit budget line at that point. This is not an accident of drawing. If an indifference curve *were* tangent to a unit-price budget line anywhere along the 45° ray, it would imply that the consumer was perfectly indifferent between future and present consumption: he would be willing to trade 1 unit of consumption today for exactly 1 unit of consumption next period. Note also that, from the construction of this family of indifference curves, a tangency of an indifference curve with a unit budget line will occur below the 45° ray: with equal “prices” for consumption now and next period, the consumer will choose to consume more than half of it in the first period. Any relative price of consumption at the two dates that would induce him to consume half or less of it in the first period would have to be steeper than budget line AA – that is, the “price” of consumption in period 1 would have to be cheaper in some sense than consumption in period zero.

Now, let’s go back to the fact that our consumer can let his present “endowment” of “stuff” grow by just letting it sit (we ignore how much he might get next period if he really bent his back to it, a more difficult problem altogether). We show in Figure 3.23 a curve that looks just like the production possibility curve of the production chapter (it is the same). The horizontal intercept of the curve at S_0^* indicates that our consumer could “eat” his entire endowment in the present period and, that if he were willing to wait, deferring all consumption until next year (it will be very convenient to consider our periods to be years), he can eat $S_0^*(1+g)$. One of the problems he faces is that if he eats everything this year, he’ll starve next year, and if he tries to wait and eat it all next year, he’ll starve this year. He’s looking for the “happy middle.”

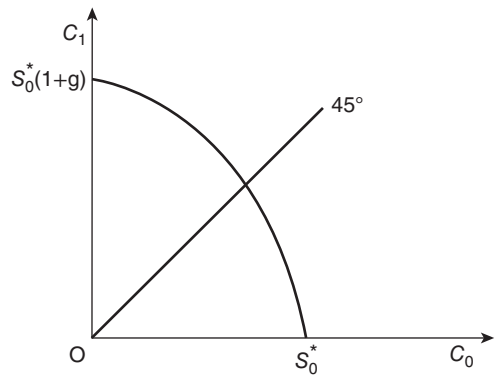


Figure 3.23 Possibilities tradeoff between present and future consumption.

The consumption good in each period is exactly the same thing, so there is no reason for its price per unit to differ, so its relative price should equal 1. The slope of the budget line (relative price line) of consumption in the two periods is the (negative of) relative price of consumption in the two periods (future over present). From Figure 3.22, we saw that some relative price greater than 1, however, was required to persuade the consumer to consume equal amounts in the two periods. The slope of the relative price line that will accomplish this is (the negative of) one plus the interest rate. That is, looking from the present toward the future, the present price of the future consumption must be less than the present price of present consumption. This is a discount that the consumer places on future consumption relative to present consumption.

We can look at this slightly differently. Give the consumer a two-period budget constraint that gives him a constant income. In period zero he receives y_0 income in terms of the consumption good. (Since both income and consumption are measured in terms of the same, real good, the implicit relative price of the good in the two periods is simply unity.) His consumption of that income we call c_0 . If he eats all of his period-zero income in that period, $c_0 = y_0$. Anything he doesn’t eat in that period, he can earn interest on (think of the uneaten part of the income simply growing) at the rate r . In the second period, he gets income y_1 . If he ate less than y_0 , he can carry over to the next period $m_0 - c_0$, on which he earns

interest at the rate r . We can write his two-period budget constraint (“intertemporal” budget constraint) as $(1+r)c_0 + c_1 = (1+r)y_0 + y_1$. He can shift income and consumption back and forth between periods. If he borrows in period zero ($c_0 > y_0$), he has to pay back $(1+r)(c_0 - y_0)$ in the next period, which means that $c_1 < y_1$.

If we divide that budget constraint by $(1+r)$, we get a slightly different perspective on these same relationships: $c_0 + c_1/(1+r) = y_0 + y_1/(1+r)$, which gives the budget constraint and consumption in terms of present values. A unit of consumption in the second period is worth less, from the perspective of the present period, than a unit of the same consumption now – less at the rate of $1/(1+r)$. Similarly with the present value of next period’s income.

Figure 3.24 shows an intertemporal budget line, with a set of incomes y_0 and y_1 , which represents the consumer’s endowment, and two indifference curves representing alternative preferences for present and future consumption. With the given endowment and the interest rate represented by the budget line, a consumer with preferences represented by I_1 will be a lender, consuming less in period zero (c_0^1) and more in period 1 (c_1^1). A consumer with a greater preference for present consumption, represented by I_2 , will be a borrower, consuming more in period zero (c_0^2) and less in period 1 (c_1^2). A sufficient increase in the interest rate, represented by a steepening of the intertemporal budget line, would turn our

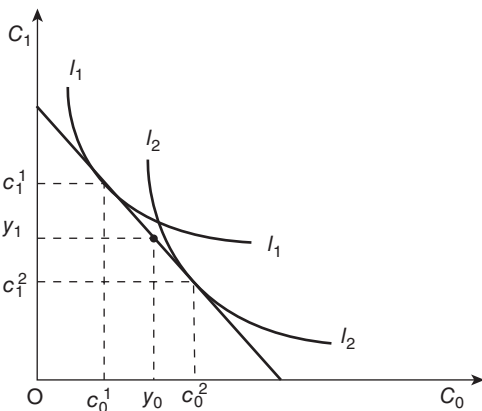


Figure 3.24 Different preferences for present and future consumption.

borrower (I_2) into a lender, but would leave our present lender (I_1) still a lender.

The concepts of discount rate and present value will be used extensively in problems that involve more than a single time period, which includes most interesting economic problems as we encounter them in the observational world. The choice of consumption quantities, c_0 and c_1 (which could, of course, be extended to $c_0, c_1, \dots, c_n$) typically is called a “consumption stream,” as the incomes in the two periods are called an “income stream.” In cases we will address later, the consumer will be able to choose, or influence, his income stream as well as his consumption stream.

3.10 Durable Goods and Discrete Choice

Having introduced the choice of consumption between time periods, we have the tools to extend the types of goods whose demand we consider beyond items that are, at least implicitly, consumed fully in the period they are purchased. Many consumer goods are intended to “last” several time periods, the consumer receiving a flow of services from the stock of equipment. This is the same stock-flow distinction we introduced in the production chapter, where we noted the difference between, for example, a person who supplies labor (the stock) and the labor services he or she provides over a number of periods. Durable goods include equipment used in production as well as consumer items intended to offer services over time, but we will concentrate on the consumer items to some extent in the present treatment. Probably the premier durable good consumed by ordinary people in the ancient Mediterranean was their housing. Among durable productive items, vehicles – including ships and boats – animals, and more complex tools are common examples.

To acquire the services of a durable good, one could either rent it period by period or acquire the good all at once (ignoring “consumer credit” for the moment). There frequently are excellent reasons for the availability of rentals to vary across types of durables, for reasons that are themselves interesting, but we will concentrate presently on

the purchases of consumer durables. The concept of user cost is important when thinking about consumer durables. A simple way to think about user cost is to compare the current purchase price with what you could get for the used durable next period. The difference is user cost. A simple expression for user cost is $p_0 - p_1(1-\delta)/(1+r)$, in which p_0 is the price the consumer must pay for the item in the present period (purchasing the entire stock of the good), p_1 is the price at which she can sell it next period, δ is the physical depreciation rate, and r is the interest rate. The second term of this expression is the present value of what can be gotten back for the used piece of equipment next period – what's left of it, that is, since the fraction δ of it has been “used up.” If the price next period is expected to be the same as the price this period, the user cost simplifies to the depreciation and an interest charge: $\delta/(1+r)$; if there is no depreciation, the user cost is just the interest charge, $1/(1+r)$. Note, however, that the user cost could be zero or negative, if the consumer gets a capital gain off the durable in the second period, which is what happens if $p_1 > p_0$. On the other hand, if the sales opportunities for the durable in the next period virtually disappear (p_1 gets very small), the durable looks very much like a nondurable consumption good (apples and bread), inasmuch as the user cost gets very close to the full, purchase-period price p_0 .

Let's consider somewhat further the relationship between the user cost for a durable good and its price. To do that, think in terms of a rental market in which the owner of the durable charges an annual rental equal to the user cost an owner would incur if he owned it himself. Then at any present period, the owner can look forward to an income stream in the form of those rentals each period: $R_0 + R_1/(1+r) + R_2/(1+r)^2 + R_3/(1+r)^3 + \dots + R_n/(1+r)^n$, in which R_t represents the rental he expects to receive in period t , and r is the discount rate, which we may suppose to represent his time preference rather than necessarily a market interest rate – although a market interest rate must reflect a combination of individual discount rates. He would be willing to sell this durable good for a price that would leave

him no worse off than holding the good (assuming no uncertainties in the rentals or the survival of the good), which would make the durable price $P \geq \sum_0^n R_t/(1+r)^t = (R_0/r)[1 - 1/(1+r)^n] = R_0/r$ for identical rentals in every period and a large number of anticipated periods (making for a large value of n). If P stays much above this value, other suppliers of the durable will undercut him ever so slightly, beginning a process that leaves all offerors of the durable offering the good for a price neither higher nor lower than R_0/r . This is quite a powerful result. If we know the sale price of a durable good and its annual (or whatever period) rental price, we can infer the prevailing discount or interest rate, even if credit markets are not particularly active and we are unable to get direct quotations on loans.

This description of the user cost of a durable abstracts from several important problems but also hints at where some others arise. First, when doing her planning about buying the durable, the consumer has to form expectations of what the next period's price will be (or if part of the plan is to use the item for several periods, then the price to be forecast is several periods ahead). Looking forward, the buyer (or potential buyer) of a consumer durable has to make a forecast of what the resale price will be, with the knowledge that that price could be higher or, more problematically, lower. This means that the user cost of a durable is not fully known until after the period of use, which puts some element of uncertainty, and hence, risk, into the consumer's calculations. Second, the market for used durables – that is, the ability to sell them – may be poor. There is an important information problem in the used (secondary) market for durable goods: the seller knows how well the item he is offering works, but the potential buyers know that some of the individual units of this particular type of durable work pretty well while some of the others are “lemons” – that is, they never did work very well, and that's why they're being unloaded. It is difficult to observe the quality of the used durable, so the buyer has to form some expectation of the likelihood that the item will “work” after he buys it. The offering price from the seller is the value of the item to the seller; if the potential

buyer were sure it'd work when he got it home, that would be his valuation of the item as well, but he recognizes that this particular item may be a lemon, so he marks down his offer price by his probability that it really is a lemon. This type of informational limitation can seriously disrupt the market for second-hand durables, because no "good" durable items will be transacted (the seller won't take less than his valuation of them), leaving only the "lemons" on the market, with a price that spirals toward zero, as buyers as a whole raise their expectation that whatever items appear on the second-hand market are of low quality. The virtual absence of a secondary market for durables will eventually affect the price of new durables.

Third, we have abstracted from any operating costs the user may incur during the period of use. Animals require fodder and water, the costs of which might change or differ from their expected values at the time of acquisition. Power sources are a prime source of operating costs for many durables, and these inputs (we use language from production theory because using a durable is indeed a production process) would have been supplied primarily by humans or animals. Producer durables (as contrasted with consumer durables) would have required some other inputs to work with, and the availability of those may have been uncertain.

Next, let's compare the aggregate demand for a durable good with the demand for one of these items by any individual. If we consider the total demand by, say 5000 people for some type of durable good, we can expect maybe several hundred of the items to be transacted during any year. This makes it clear that we are talking about something on the order of one-tenth or one-twelfth, say, of a unit of one of these items per person. How can a person have a demand for one-tenth of, say a wagon? Clearly, at the level of the individual consumer, we have a different type of problem than we had with the demand for apples and bread. We were able to consider apples and bread as pretty much "continuous variables" – the consumer's demand for them could have been 1 or 2, or 15 or 16 or 17, and so forth – not zero or one as it is with the typical

durable good. This brings us to the problem that has been called "discrete choice." The theory of discrete choice has taken the route of considering those demands as probabilities of purchasing said durable good during any period, where the probability of purchase is more or less the usual function of relative prices (user cost) and income, as well as other characteristics of the unit considering the purchase. If a house were the durable under consideration, the consuming unit would be a household (or family), and the characteristics of the consuming unit that would be relevant to this demand would include the number of members, their ages, activities to be undertaken in the house (production as well as eating, sleeping, leisure, and so forth).

We can still use utility theory and indifference curves to study discrete choices. Figure 3.25 shows the indifference curves of two consumers who contemplate the purchase of a consumer durable (D) alongside their usual purchases of a "continuous" consumer good, q . Let's also study this problem as an expenditure-minimization task for these two consumers: that is, they each minimize the expenditure required to reach a given level of utility that their household resources let them reach, given the prevailing structure of prices that they face. We can write this problem as $\text{Min } pq + cD = y$, subject to $U = f(q, D, \Theta)$ where p is the price of the "continuous" consumer good, q is the quantity of that

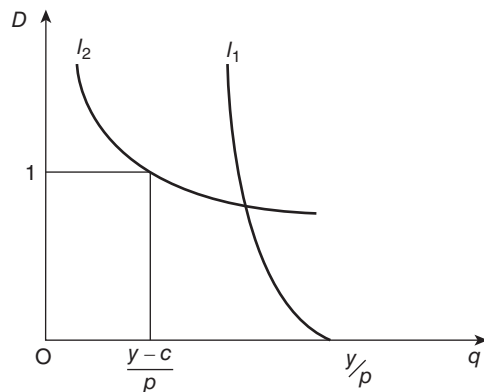


Figure 3.25 Discrete choices.

good, c is the user cost of the durable, D is the stock of the durable (which can take the value 1 or 0), y is household income, and Θ is a set of household characteristics. We have constructed Figure 3.25 so that one of these consumer units will purchase the durable (group 2) and the other will not (group 1). For the consumer group that doesn't purchase the durable, expenditures are simply pq , so we can rewrite the expression for their income as y/p . Indifference curve I_1 cuts the horizontal axis (where $D = 0$) at the value of q equal to total income, y/p : they spend all their income on the consumer good. The other household, which purchases a unit of the durable, buys $(y-c)p$ of the consumer good (rearranging the expenditure expression, letting the value of D be 1). The slope of indifference curve I_2 at $(1, y-c/p)$ is the relative price of the consumer good and the durable, and it is clear from the shape of indifference curve I_1 that the other household will never find a marginal rate of substitution (MRS) between consumer goods and the durable equal to that relative price for a positive quantity of the durable. What makes for such different shapes of indifference curves? We know that the utility function incorporates tastes, but we should not feel comfortable assigning such a major, economic result as whether or not to purchase a durable good to tastes alone. There may be income differences between the consumers represented by I_1 and I_2 , but differences in income should mostly just shift indifference curves toward or away from the origin of the diagram, not change the shape of the curve: remember the requirement that indifference curves be parallel. This leaves the variable "other household characteristics" responsible for the difference in the shape of these two indifference curves, the immediate implication being that characteristics of the consumer unit (individual or family) can affect the MRS between goods. The more far-reaching implication is that, if this is not simply another term for differing tastes, then some important production effects are contained within the variable we have just introduced as "other household characteristics." Pursuing this observation will take us into what is called household production theory.

We can examine this problem slightly differently using the same utility apparatus. Take the general form of the utility function for either

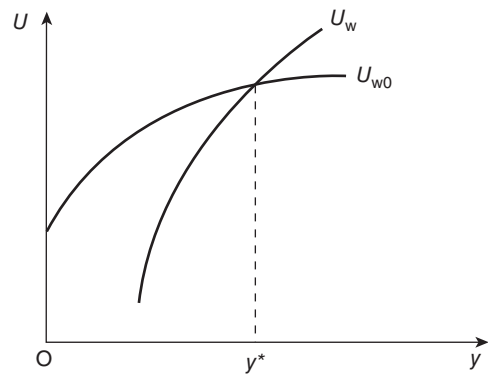


Figure 3.26 Effect of income on allocation of expenditure between nondurables and durables.

of our consumer households, $U = f(q, D, \Theta)$, and let one of the consumers (it doesn't matter which one, as long as it is the same one) compare its utility with and without one unit of the durable: $U_{wo} = f(y/p, 0, \Theta)$ without the durable and $U_w = f((y-c)/p, 1, \Theta)$ with it. Whichever choice gives the higher level of utility will be the consumer's preferred alternative. If $U_w > U_{wo}$, the durable will be bought. What would make the utility of either choice higher? Income is capable of doing the trick, as Figure 3.26, which plots the value of utility against income, shows. The U_{wo} curve is higher than the U_w curve at low income levels because allocating all income to the continuous nondurable good purchases a considerable quantity of it, while the U_w curve shows the effects of low consumption of the nondurable at low income levels. However, as income increases, allocating all of it to the nondurable eventually encounters severely diminishing marginal utility, flattening out the utility curve at higher levels of income. As income rises, the utility derived from allocating some expenditure to the durable good rises sharply over the same range that sees a flattening of the nondurables-only curve. At the level of income designated by y^* , the consumer would obtain more utility by buying one unit of the durable.

This example of the single household's demand for a single unit of a good can be expanded in two directions. First, in many cases, the purchase of a durable good is a replacement for one the consumer already owns. Second, the choice the consumer really faces may not simply be to buy

a durable good or not, but rather which durable good to buy. The first issue takes us in the direction of the timing of durables purchases, and the second into issues of variety and differential quality. The services provided by an existing durable good may get smaller over time as depreciation removes some of the effective stock of the good, but options may exist for maintenance to temporarily restore some of the service flow. The purchase price of the minimum size of a durable may be considerable, relative to the prices of consumer goods and even relative to current-period income. Uncertainty about future income and future interest rates, or dips in current income and blips in current interest rates, combined with the ability to coax some extra service flow out of an existing stock, can lead consumers to defer purchases of new durables. When many individuals' incomes are affected by some common forces or events and an interest rate is faced by all consumers, the aggregate demand for any individual durable, and possibly the entire array of durables, can fluctuate more than the aggregate demands for consumer goods. Under many conditions, the deferral of durables purchases can contribute to business cycle fluctuations. Even in the economies of ancient Egypt and Mesopotamia, if a sizeable proportion of labor time were spent in providing durable goods – housing being a major component – events that caused consumers to defer durables purchases could have led to the calling of loans, idling of labor resources, and the usual accumulation of economic misfortunes associated with downturns in the business cycle.

Another issue in the aggregate demand for durables is the length of time required for a change in demand to be supplied. If some exogenous change occurred that prompted consumers to want to hold a larger stock of a particular durable good – for example, the prospects for future income brightened considerably, interest rates fell, part of a major city burned or was destroyed by earthquake – it is unlikely that this new supply would be furnished immediately. More rapid production implies working up the supply curves of the firms or individuals producing the good in any particular period, which would drive up the price. Instead, the stock of the durable actually held by consumers is likely to adjust more slowly, taking several periods to reach a new level of desired stocks. This is

another difference between durable goods and consumer goods.

The subjects of variety and quality take us into a separate topic which has application to consumer goods as well as to durables. Before we proceed to those topics, we pause to point out one more, far-reaching aspect of durable goods. We have pointed out that durables are demanded for the flow of services they are expected to yield over time. This amounts to planning one's consumption of the durable over time, but consumers' demands for any particular good are influenced by their demands for many other goods – think back to the cross-price elasticities of demand. When the demand for one good changes, those cross-elasticities transmit this information to other items in a consumer's consumption set. The plans for the future consumption of durables implicitly contain at least partial plans for other consumption – of nondurables as well as other durables. The interest rate, as an implicit valuation of future consumption relative to present consumption, is a direct mechanism for linking present and future decisions. This is looking forward. Looking backwards, previous decisions about durables purchases affect current decisions about how to allocate other resources, and past, present, and future resource allocation decisions are tied together inextricably.

3.11 Variety and Differentiated Goods

For ordinary utility and demand theory to be particularly useful for understanding the demand and supply dynamics of many durable goods, we must face directly the fact that many of those goods are differentiated from one another. Each good may be thought of as consisting of many features or characteristics that can be combined in many different ways – size, quality, and so forth. Having made the observation for durables, we easily can extend the characterization to many nondurable consumer goods. Differentiation and variety are ubiquitous.

Several principal, complementary approaches to the study of product variety and differentiation have emerged. The first approach can be called the hedonic good, or hedonic price, model. The second, called the characteristics model, focuses

on the supply of and demand for the characteristics embodied in goods. The third approach addresses product variety from the perspective of the industry structure called monopolistic competition (discussed in Chapter 4). The three approaches are interrelated, the first two through their focus on product characteristics; the last two through their attention to the concept of optimal product variety provided by a monopolistically competitive industry.

The theory of hedonic prices has followed behind the empirical application of the concept, which has striven to attribute the aggregate prices of durable goods such as houses, automobiles, and household appliances to “partial” prices of their characteristics, such as number of rooms, square footage, and appliances in the case of houses; horsepower, number of doors, weight, and so forth, for automobiles; and size and user cost for appliances. The product characteristics are identified in quantifiable forms so that the “price” of one unit of the characteristic can be estimated; the sum of the characteristic prices times the number of units of each characteristic equals the aggregate durable good price, less a statistical error term. For example, a hedonic housing price equation might look like $P_h = a + p_{\text{room}} \# \text{ of Rooms} + p_{\text{sq ft}} \text{ Square feet} + p_{\text{heating}} \text{ Size of heating/cooling unit} + e$, where the a coefficient might be assigned the value of zero or might be allowed to take a nonzero value to account for some fixed costs or costs of omitted characteristics; the p_i are prices of the characteristics, times which their quantities are multiplied; and the e term is a statistical error from the estimation. The statistical estimation (via the technique known as multiple regression) estimates the a and p_i terms, using the error term as part of the estimation procedure but not returning a summary estimate associated with it but rather an array of errors specific to each observation in the sample.⁹ Hedonic price estimation has been conducted for nondurables ranging from different qualities of food grains to wines to “off-market” crude oils, so the technique is not restricted at all to durables.

While the hedonic price model has relied on the concept of quantifiable characteristics of goods, the commodity characteristics model of Kelvin Lancaster has begun with the concept of the demand for and supply of characteristics.

In the characteristics model, consumers actually begin with utility from characteristics that just happen to be embodied in goods. The characteristics model began, and primarily remains, as a theoretical effort, most of the empirical research remaining within the hedonic price framework. Nonetheless, the conceptual explorations with the characteristics model offer some useful insights into both the demand for differentiated goods and their supply.

The characteristics model defines a product “group” as a group of goods and potential goods (goods that might be produced under some circumstances but not currently produced) that possess the same quantifiable characteristics, the differences among the products being identified as the quantitative differences in their characteristics’ contents. Product differentiation is equivalent to variations in the characteristics’ contents within groups of goods. For any particular product in a group, there are cost tradeoffs among the ratios of characteristics; for example, at a constant unit production cost of a good containing characteristics A and B , the quantity of characteristic A that can be embodied in the good can be increased only at the cost of giving up some quantity of characteristic B . This tradeoff between the quantities of different characteristics is imposed by production technology, and it ultimately places limits on the number of different product varieties that any group of demanders can pay for. Additionally, there are tradeoffs between the additional benefit (utility) that consumers derive from increased product variety and the scale economies (possibly deriving from fixed costs rather than from long-run increasing returns to scale in production functions) foregone by producing smaller numbers each of a larger number of varieties.

In the characteristics model, not every variety that the economy has the technological knowhow to produce will be produced. Depending on the number of consumers contributing to the demand for these products, the substitutability among different varieties of products within the same product group, and the economies of scale in producing the different varieties, a larger or smaller number of them will be produced. The principal tradeoff for the consumer is the increased utility derived from varieties closer to

her taste versus the lower cost of standardized commodities for which greater scale economies are available.

Technological changes that gave greater scale economies at smaller levels of output would encourage a greater number of varieties. Otherwise, the absence of scale economies would foster the proliferation of variety. Lancaster himself speculates about the relatively limited varieties of goods in production in the preindustrial period ("before the development of machine-based manufacturing"), suggesting that a situation of few apparent scale economies might have been combined with highly constraining fixed costs in the form of craft skills which were costly to replicate (Lancaster 1979, 11–12). While it might appear that the combinations of materials, simple tools, and labor could have been subdivided into the production of a large number of varieties of goods, in fact the labor skills embodied in the craftsmen would have represented a type of scale economy that was, in many cases, able to produce a small number of nonuniform, and hence high-cost, products for wealthy clients by exceptional exercise of skill. For the vast majority of other customers, considerable uniformity of products (low product differentiation) would be the consequence of the scale economies of skill in executing that limited range of products. This appears to have the seeds of several hypotheses that might be confronted with ancient data.

The optimum product variety model places an industry that provides differentiated goods within the context of the rest of the economy that provides homogeneous goods and asks, among other questions, what the optimal number of varieties provided by this industry is, and what factors influence that number. In an influential presentation of this model, Dixit and Stiglitz noted that the ordinary utility function defined in homogeneous goods contains the strong hint that variety itself is valued. Consider an indifference curve drawn for two goods, call them *A* and *B*. From the convex shape of the indifference curve (bowed in toward the origin of the graph), we can see that a consumer who was indifferent to (*A*, *B*) combinations of (0,1) and (1,0) would prefer ($\frac{1}{2}$, $\frac{1}{2}$) to either specialization (Dixit and Stiglitz 1977, 297). This insight leads to the specification of a utility function in which the quantity of homogeneous goods still contributes to utility,

but also the *number* of varieties of differentiated products contributes to utility.

We can treat the combination of homogeneous and differentiated goods with a two-tiered utility function, $U = U[u_1(\blacksquare), u_2(\blacksquare), \dots, u_n(\blacksquare)]$, in which the u_i are "subutilities" derived from the consumption of products *i* and the $U[\blacksquare]$ is the upper tier utility function that converts all the subutilities into an overall utility measure. If some product *i* is a homogeneous good, only the quantity consumed of it affects the consumer's subutility: $u_i(q_i) = q_i$. But if the product is a differentiated good, then $u_i(\blacksquare)$ depends on the quantity of each variety consumed: $u_i(q_{i1}, q_{i2}, \dots, q_{in}) \equiv (\sum_j q_{ij}^{\beta_i})^{1/\beta_i}$, in which $\beta_i = 1 - 1/\sigma_i$ and $\sigma_i > 1$ is the elasticity of substitution between varieties, which is greater than 1 to make sense of the type of competition that exists between the differentiated products.¹⁰ Next, if we let E_i be the expenditure allocated to product *i*, the level of subutility achieved from this expenditure is $u_i = n_i^{1/(\sigma_i-1)} E_i/p_i$, where n_i is the number of varieties of differentiated good *i*. Then the consumer's choice problem can be simplified to the following:

$$\text{Maximize}_{(q_1, q_2, \dots, q_n)} U[n_1^{1/\beta_1}, n_2^{1/\beta_2}, \dots, n_n^{1/\beta_n}]$$

subject to $\sum_i p_i n_i q_{fi} \leq E$, where E is the aggregate spending level. This expression says simply that the consumer maximizes her utility, which is represented by all the relationships inside the $U[\blacksquare]$ term, by adjusting her choices of the quantities of each variety of each good consumed; and that she does this while keeping her total spending down to at least the total income she has available to spend. (A restriction of this particular version of the variety model is that the quantities of each variety consumed be equal; this considerably simplifies the calculations and doesn't, in the end, actually subtract that much from the analytical insights available from the model.) Remember that a demand function can be derived by rearranging the first-order conditions of such a utility-maximization problem. The demand function that comes out of this problem has a form in which the effects of prices and variety availability can be separated from those of income (just an analytical convenience, not some overwhelming insight into how the world works): $q_i^d = \varphi_i(\mathbf{p}, \mathbf{n})E$, in which the bolded $\mathbf{p}$ and

n represent the entire array of prices and varieties for all the goods that enter the utility function; the substitution elasticities are contained in the explicit form of the function φ . Rearranging this particular demand function implies that the share of spending allocated to any differentiated good i , $\alpha_i(p, n)$, is simply $p_i \varphi_i(p, n)$, which says that the expenditure share on any one of these differentiated goods depends only on the number of varieties of it that are available, their prices, and the elasticities of substitution among them (buried in φ).

This model shows that (and how) consumers value the availability of variety as well as that of quantity. Although this particular implementation of the variety model doesn't show a tradeoff between the variety and quantity, it is only a step (if an algebraically messy one) to that result. Accordingly, the puzzled notion that occasionally appears in archaeological ruminations about imports found at a site – “This good was available locally. Why did these people bother to import some from places X, Y, and Z?” – can be resolved easily. People, then as now, probably liked variety (disliked boredom in their consumption) and were willing to pay for it. Why did a land like Cyprus, a prodigious olive-oil producer, apparently import some oil from Crete and probably mainland Greece during the Late Bronze Age? A case of coals to Newcastle? The simplest explanation is that there were a number of distinct varieties of olive oil, either useful in different roles or appealing to tastes or incomes of different people. Cyprus, Greece, and Crete each produced some different varieties, some of which probably didn't get traded because the identical oil was produced in the other region, and some of which did get traded because they weren't.

A key idea behind the variety concept is that, for any given product, as a consumer consumes more of it, her marginal utility falls. If there were something else to do with the good besides consume it that yielded utility to the consumer – say give it away to somebody else and let them consume it – when the marginal utility of consumption falls to the level at which the alternative use of the good is competitive with direct consumption, additional units of the good go into the alternative use. If there are only a few varieties of consumption goods, declining marginal utility sets in fairly soon, and rather than get bloated

on continued consumption, our consumer starts giving units of these few goods away – possibly as religious dedications. Now, suppose that a number of new varieties of consumption goods are introduced. For each good, declining marginal utility is still the rule, but now income is spread over consumption of more varieties, so marginal utility for any one good will not decline as quickly with income as it did with fewer varieties, because the consumer can switch from consumption of one variety to consumption of another variety. As a consequence, the goods may have to queue up farther back in the line. The implication for ancient religious behavior is that the introduction of more varieties of consumption goods could have had a dampening effect on the magnitude of offerings of all kinds to religious institutions, be they temples, or neighborhood altars, or a priestly establishment. Whether sufficient product variety developed during antiquity for this effect to have shown up is an empirical question. If we find periods (and places) of declining religious offerings without indications of unambiguous income decline that could account for a depression in offerings, the variety of consumer goods could be a reasonable place to look for an explanation. Whether the variety of consumer goods increased at times and places in antiquity sufficiently to generate such an effect is, of course, an empirical question. And of course whether sheer, common greed by itself is sufficient to loosen the attachment to one's deities required to stiff them in offerings can be made an empirical rather than a purely philosophical question.

3.12 Value of Time and Household Production

The basic theory of consumption places goods purchased or otherwise produced directly into the utility function, as if simply having them could produce utility. It takes time to actually consume virtually all goods: having acquired (bought or grown) beans, it takes time to prepare a meal, then further time to eat it; without eating the meal, it is difficult to see how it could contribute to one's feeling of wellbeing. Gary Becker's (1965) theory of the allocation of time (also called the household production model) has both formalized much prior common sense and

revolutionized the way economists think about issues throughout the discipline.¹¹

Recall the formulation of a utility function as we have used it so far: $U = f(X_1, X_2, \dots, X_n)$, in which the X_i are goods numbered 1 through n . The household production model starts by revising the formulation of what is actually consumed. Thinking in terms of the meal prepared and eaten rather than simply the food items used, replace the goods in the old utility function with final consumption goods, Z_i : $U = f(Z_1, Z_2, \dots, Z_n)$. These final consumption goods are themselves produced by the consumers with combinations of “market goods” (the “market” part really is unnecessary except as a means of distinguishing the initially acquired commodity from the final consumption act; these goods could be produced in one’s garden or field, bartered for with neighbors, or ordered out of a mail-order catalogue). We represent this formulation as a relationship between inputs of time and market goods: $Z_i = g_i(X_i, T_i)$. We subscript the functional notation of the production function (g_i) to recognize that each final consumption good Z_i is produced with a different production function. The expression says that each final consumption good is produced with some combination of a market good X_i and consumption time T_i . (We could think of the production of some Z_i as involving several market goods, but for simplicity of exposition we use just one.)

The consumer maximized the old utility function subject to an income constraint that said, for all practical purposes, that he couldn’t spend what he didn’t have – his expenditures equaled his income from labor earnings and any assets that might yield income each period: $\sum_i p_i X_i = Y + A$, in which the p_i are the prices of the goods X_i , Y is labor income, and A is unearned, or asset, income. Labor income can be defined as the wage rate times the hours spent working (it is not necessary to think in terms of a labor market to arrive at this specification: the wage is simply the value of marginal product of labor in whatever activity or activities the consumer engages): $Y = wT_w$, where T_w is the amount of time spent working outside the home. The total time available, T , can be divided into the amount spent working and the amount spent in consumption activities: $T = T_w + T_c$. We have already identified the amount of time spent in

producing a final consumption good as T_i , so the sum of all the times spent in producing all one’s final consumption goods equals the total time spent in consumption: $T_c = \sum_i T_i$.

We now come to a rather subtle, but very important step. We substitute T_w out of the new budget (income) constraint, recognizing that $T_w = T - \sum_i T_i$. The budget constraint now reads: $\sum_i p_i X_i + \sum_i T_i w = wT + A$. This has the effect of saying that the “full” income constraint of the consumer includes earnings (wage) income equal to what he could earn if he literally worked all the time – including sleeping time! (Sleep simply becomes one of our home-produced consumption goods, and its cost in terms of foregone “other things” that one could produce with the time will be one determinant of how much one consumes of it.) This result has emerged not through any special assumptions or sleight of hand, but from the simple division of a person’s time into two different and exclusive types of use in what amounts to nothing more than a codification of common-sense observation.

For some temporary expositional simplicity, let’s assume that the production functions for the Z_i use fixed proportions of market goods and consumption time. In the production chapter we characterized this type of production function diagrammatically as having L-shaped isoquants, but we did not show the formula for such production. We’ll show it here: $T_i = t_i Z_i$ and $X_i = b_i Z_i$, which says that there are fixed input-output coefficients t_i and b_i for time and market goods in the production of a unit of each final consumption good. Expressed alternatively, $t_i = T_i/Z_i$ and $b_i = X_i/Z_i$, which simply says that the input-output coefficient t_i is the amount of consumption time that goes into a single unit of the final consumption good, and similarly for b_i . Now we can substitute this relationship between the Z_i and the two inputs (market goods and consumption time) back into the full income constraint to get $\sum_i (p_i b_i + t_i w) Z_i = wT + A$, which expresses the constraint on final consumption in terms of final consumption. This expression also gives us a formula for the cost (or “shadow price,” to which we have referred earlier) of final consumption in terms of “market costs” and time costs. If we define $p_i b_i + t_i w = \pi_i$, π_i is the shadow price of final consumption, or the full cost (price) of a unit of final consumption.

The consumer's maximization effort in the case of the simple utility function led to the efficiency condition that the ratio of marginal utility to commodity price was the same for all goods: $MU_1/p_1 = MU_2/p_2 = \dots = MU_n/p_n$. The home-producer's utility maximization yields comparable conditions, except that the shadow prices of final consumption replace the prices of directly purchased commodities: $MU_1/\pi_1 = MU_2/\pi_2 = \dots = MU_k/\pi_k$. In addition, the usual conditions for efficiency of production emerge from the same utility maximization: $P_{Z_1}(T_c)/w = MP_{Z_1}(X_1)/p_1 = MP_{Z_1}(X_2)/p_2 = \dots = MP_{Z_1}(X_n)/p_n = MP_{Z_2}(T_c)/w = MP_{Z_2}(T_c)/p_1 = MP_{Z_2}(X_1)/p_1 = MP_{Z_2}(X_2)/p_2 = \dots = MP_{Z_2}(X_n)/p_n \dots MP(T_w)/w$. The ratio of marginal productivity to purchase price of each purchased good is equalized across all uses of all purchased inputs. Each of these ratios is equalized to the ratio of marginal product of time to the opportunity cost of time (which we have identified with the wage rate for simplicity). And these latter ratios equal the ratio of marginal productivity of labor in external production.

One might well ask how an ordinary family, not to mention one composed of three generations and possibly aunts, uncles and cousins, could ever reach such a pinnacle of optimized bliss. Fair question. First, note that these conditions do not require anyone to know average productivities or to make agreements or concessions regarding that statistic. The marginal values require only the immediate observation of how much of whatever intended outcome derives from the last unit of effort, in the case of time, and how much the last cup, or variation in cup size, contributes to some final consumption activity that uses cups as one of its purchased inputs.¹² Even such observational efforts are costly in terms of time and possibly the distaste of remembering and calculating. Consequently, in situations in which time has low opportunity costs (that is, when labor productivity is low, w is low), time probably will not be accounted too closely. Come harvest season for agricultural societies, however, time will be accounted more carefully because the demand for it ("things to do") outstrips its supply (everybody else is busy too). A society with low labor productivity may account the marginal contributions and marginal costs of its "purchased" inputs more

carefully than its time. But this is nothing more than saying that the marginal benefits of close accounting of goods and time vary predictably across circumstances, and that the activity we call calculation will be conducted to the point where its marginal benefit equals its marginal cost. How will our consumer know when this point has been reached? "Man, this is getting to be a drag!"

The exposition so far involves some restrictive assumptions that can be relaxed at the cost of some complication: that the wage is constant across units of time, that all the household production functions have constant returns to scale, and that the input-output coefficients are constant for given p_i and w . For present purposes we need not show the way out of these specification constraints; the reader wanting to know that can consult the original publication.

Consider some of the implications of this characterization of consumption. First, part of the cost of any consumption is the foregoing of "money" income (or income in any other numeraire) that could be earned with the time devoted to the consumption. As any reader of this discussion can appreciate, the cost of tuition, books, and room and board are only a portion of the cost of education: the years taken out of earning time, or used in lower-paying activities, may constitute a larger proportion of the cost of education than the "out-of-pocket" expenses. This cost is larger for so-called "returning" students, who can earn more in the labor market than 18- and 19-year-olds.

Second, the source of income makes a difference in the effects of income changes. A change in what we have called asset income, for want of a better term, will have different effects on consumption patterns than a change in labor income. Suppose labor income rises; the cost of spending time making consumption goods increases but the cost of "market goods" (we could think of them as purchased or even bartered goods) remains the same. With an increase in asset income, neither the cost of consumption time nor market goods changes. Consequently, if income from labor earnings rises, the relative prices of time-intensive final consumption goods will rise and the consumer can be expected to substitute away from them. If we allow for substitutability between time and market goods

in the production of final consumption, people experiencing a rise in labor income will produce the same consumption goods with a higher proportion of market goods to time. These are just the substitution effects. Changes in income from both sources also have income effects, which will increase the demands for all goods, which will increase the demands for market goods as well as consumption time. Following a change in earned labor income, the income and substitution effects on the demand for market goods will reinforce each other, possibly inducing the consumer to divert time, on balance, from home production to production outside the home. The substitution effect from an increase in asset income could be very weak, although to the extent that the income elasticities of demand for final consumption goods differ, there will be a reshuffling of both market goods and consumption time among those goods.

Related to the second set of implications, the composition of household income will affect the cost of consumption as well as the pattern of demand. Changes in the external income of any family member can induce rearrangements of the productive patterns of every other family member. Activities such as procreation and bringing up children affect the cost structure of the household as well as its demand composition; correspondingly, different families will face different costs of having children, depending on their external labor opportunities and skills. Skills and productivities change over the life cycle, and the household production model correspondingly provides the basis for predictions about life-cycle consumption and production behavior.

Third, consider the consequences of technological changes on labor productivity in and out of the home. If technological changes affect home and external labor productivity identically, both w and t_1 , the labor input-output coefficients, will change in opposite directions and will tend to cancel each other's substitution effects. However, the income effect will increase the demand for market goods to accompany the greater productivity of consumption labor. If a technological improvement affects the productivity only of consumption time, the substitution effect of the change in relative input costs will cause the demands for time-intensive consumption goods to rise. Marginal labor productivity at home will

rise relative to what it is externally, inducing a transfer of labor from external production to the home. Compare that with the consequences of a technological improvement that affected labor productivity only outside the home. The rise in w would raise both "money" income and the relative cost of time-intensive consumption goods. The substitution effect would raise the demand for market goods as consumption time was transferred to external production; the increased demand for market goods would reinforce the transfer of labor from home to external production.

Fourth, the cost composition of consumption goods will affect demand elasticities for market goods. The same percentage increase in the price of a market good whose cost share in home production is, say, 60% will create a larger percentage change in the cost of the final consumption good than it would if the market good accounted for only, say, 10% of the total cost. Similarly, changes in market good prices will have larger impacts on families with low external labor productivity (w) than on those with higher external productivity; that relationship can be reversed according to differences in home labor productivity.

Fifth and finally, the household production models shed new light on two longstanding concepts, leisure and tastes. While leisure commonly is thought of as doing nothing particularly productive, the insight that emerges from this model is that whatever we do is costly, and that we frequently combine purchased inputs with so-called leisure time, making the consumption of leisure even more costly. As Becker himself has noted, only pure contemplation is a household production activity that uses no purchased input – but even then we might have to do that naked in the forest and exclude the transportation costs to the forest. The concept of leisure as everything other than "work," is not particularly useful analytically. Regarding tastes, Becker and George Stigler have developed the case that much of what is called tastes is in fact education that increases productivity in specific areas of consumption (Stigler and Becker 1977). The education could be called skills, the acquisition of which is costly and is itself produced over long periods of time, as an investment; and the model does give an explanation for changes of tastes over time.

3.13 Risk, Risk Aversion, and Expected Utility

A good deal of everyday life involves risk, even minuscule risks of death. Will the fruit I buy at the grocery store be any good? Will a new car I want to buy be trouble-free? I've survived 20 years of driving on this urban freeway; will today be the day I don't? Will my house burn down in the coming year? We will introduce some new concepts and new tools to study these types of issues systematically.

First, the term "risk" is used colloquially to refer to things that might turn out one way or another, frequently with some outcomes being thought of as preferable in some sense. Let's rename "one way or the other" "states of the world." In one state of the world, the orange I'm contemplating buying will be excellent. In another state of the world, it will look great on the outside but be terrible within. Before I buy it and peel it, I don't know which state of the world will happen. This example may seem a bit contrived, so let's consider another one with a more dramatic future. I take reasonably good care of my house, but there is some chance it will catch fire and burn down. Looking forward over the next year, the world could turn out differently. In one unfolding, my house could survive untouched by flames at the end of the coming year. In another, it could have burned down. Presently I don't know which of these states of the world will have materialized. These events are contingencies, uncertainties, unknowns. Yet we might know something about them already. Actuaries can give us excellent estimates of the likelihood of a particular type of house, given its location, age, construction materials, and so forth, burning down in any given period of time, commonly one year. The likelihoods are called probabilities.

In each of these cases, we will obtain some level of utility under either possible outcome. In the example of the house, we could construct for ourselves a utility function that was a function only of the value of the house (we've already established the fact that the flow of services we get from a durable good is related closely to the size of the stock from which those flows come in each period; and that the price of the stock

will reflect its size). If the house survives the year unburned, we would enjoy $U = U(V)$, where V is the value of the house. If it burns we get $U = U(0)$ (we could assume either that the house burns down on the first day of the year or, if it burns down later in the year it took all the pleasure out of the first part of the year; it doesn't really matter). So we know that our utility for the year will be either $U(V)$ or $U(0)$. We have constructed these two alternative, contingent states of the world so that they are mutually exclusive (they can't both happen) and complete (they include all the possibilities). If we wanted to contemplate only two rooms of our house burning, we would need to respecify the problem to include three possibilities: not burning at all, only two rooms burning, and the whole thing burning. What do we really think will happen? We could contact the actuary and find out the probability of our house burning, but how would we use that information?

Suppose the actuary tells us that the likelihood of our house burning is p , some probability between 0 and 1 in value (it can include the values of 0 and 1 as well as all the values in between). Then the probability of the house not burning is $1-p$. We can look forward to two probabilistic utilities: $pU(V=0)$ and $(1-p)U(V=V)$. The average, or in the language of probability and statistics, the expected value of our utility is just the sum of these two expected utilities: $EU = pU(0) + (1-p)U(V)$, where EU stands for "expected utility," and the utility function itself is known as the von Neumann–Morgenstern utility function after the mathematician and economist who formalized the concept (von Neumann and Morgenstern 1947, Chapter 1). More generally, we could write the expected utility function as $U(P_i) = pU(A_1) + (1-p)U(A_2)$, where the $U(P)$ is to be read "the utility of prospect i ," and A_1 and A_2 are alternative states of the world one and two. It is simple to consider a larger number of alternative outcomes: $U(P_i) = \sum p_i U(A_i)$, in which the sum of the probabilities ($\sum p_i$) equals one.

The expected value of the utility in the case of our house either burning or not burning is $pU(0) + (1-p)U(V)$, but the *utility of the expected value* of our house is $U[p \cdot 0 + (1-p) \cdot V]$, which

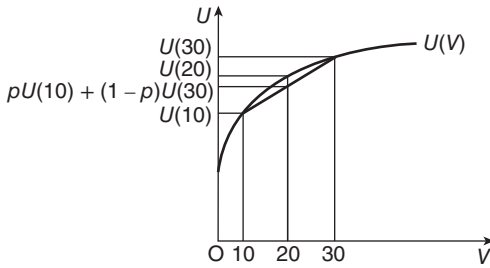


Figure 3.27 Expected utility.

need not equal the *expected value of the utility*. We show this difference graphically in Figure 3.27, in which we let V represent the value of the house. Rather than compare the value of zero with a nonzero value, lest someone think there is something special in a zero value, we consider the possibilities that the house value can take on values of *either* 10 or 30. Without intending to suggest that we live in an especially risky house, we have chosen a large probability for the eventuality that it will burn in the next year. Values along the $U(V)$ curve identify the level of utility we would get if we actually obtained values of V equal to the numbers along it. But in our contingent consumption situation, we know that our house is worth $V = 30$ if it doesn't burn and $V = 10$ if it does. An actual $V = 20$ won't occur, although 20 will be the expected value of the house. The utility of the expected value of 20 is greater than the expected value, weighted by the probabilities of occurrence, of the two events, one of which we know *will* happen. While we could get a utility of "house value = 30," we also could get utility of "house value = 10." The consumer described by this particular utility function would get a greater sense of satisfaction out of "sawing off" part of his house and having that for certain rather than taking a chance on having either amounts 30 or 10. We have an example of a person who is averse to risk – who would rather have a smaller amount of a "sure thing" than the possibility of two uncertain things that give an expected value equal to the "sure thing." Figure 3.28 shows a consumer who would gladly accept the risk involved to preserve the possibility of obtaining the higher utility. The utility curve of a perfectly risk-neutral consumer would be a

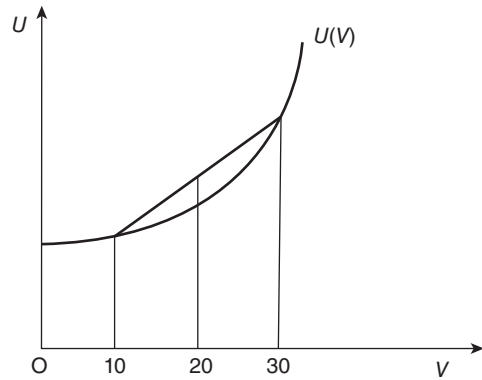


Figure 3.28 Utility curve of a risk-prefering person.

straight line in the independent variable (house value in this case). Most people tend to be risk averters in most of their decisions, although that does not preclude their accepting risks and occasionally having a "stretch of utility function" over some independent variable that leads them to favor risk. A commonly used measure of risk aversion is the quantity $-\Delta(\Delta U)/\Delta U$, which is a measure of the curvature of the utility function divided by a measure of its slope, so it could be expected to vary across the utility function. If this measure is negative, the person whose utility function is so represented is a risk averter, at least at the point for which the measure is taken.

Having introduced the concepts of risk and contingent consumption, it is a short step to consideration of insurance against the risk. This changes the *prospect* from the all-or-nothing comparison we considered in the case of the house burning to one that is "not quite all" versus "considerably more than nothing." Suppose we could purchase insurance that would replace the entire value of the house if it burned down but paid nothing if the house didn't burn. If the house fails to burn in the coming year, we have the value of the house minus the insurance payment: $V - I$, where I is the amount we pay for insurance. If the house does burn, we have 0 for the house plus V , which is replaced by the insurance, minus the cost of the insurance, I . So if the house doesn't burn, we have $V - I$ to insert in our utility

function, and if it does burn, we also have $V - I$. (This assumes quite a bit about the insurance policy we buy: it has no deductible and offers full replacement.)

Let's try a numerical example of how much a person with a utility function of the form $U = V^\alpha$ would be willing to spend on insurance. Suppose the value of the house unburned is 100 000 (we deliberately ignore the units; they could be dollars or ducks). If the house burns, the foundations still will be worth 10 000. There's a one-in-a-hundred chance that the house will burn this year (probability = 0.01). The consumer can spend some amount I on insurance; how much will I be? Make the following comparison: $U(100\,000 - I)^\alpha = 0.01 U(100\,000)^\alpha + 0.99 \cdot U(10\,000)^\alpha$. To show the effect of increasing risk aversion, we will try out two different values of α : 0.75, which is closer to linear (which would be represented by $\alpha = 1$), and 0.5. First, we have $U(100\,000 - I)^{0.75} = 0.01 \cdot U(100\,000)^{0.75} + 0.99 \cdot U(10\,000)^{0.75}$. Solving this for the value of insurance I , we get $I = 4355$. The expected value of the loss (1% of 100 000) is 1 000. So-called "fair" insurance would amount to only the expected value of the loss; the fact that this consumer is willing to pay 4355 indicates his risk aversion. Now, substitute $\alpha = 0.5$ into that formulation and find that it would amount to 9400, with the same expected loss of 1000. The greater risk premium reflects the greater degree of risk aversion the consumer would have with the lower coefficient on the house value in his utility function. The shape of the utility curve in this example would be represented by Figure 3.27; with $\alpha = 0.75$, its curvature (not simply its slope) would be flatter than with $\alpha = 0.5$. The vertical distance between the utility curve itself and the chord drawn between $V = 10$ and $V = 30$ (the straight line) in that figure represents the risk premium the consumer is willing to pay for the certain value of the associated value of V (read down to the horizontal axis) rather than the expected value. This risk premium is an amount the consumer would be *willing* to pay to avoid the risk, not necessarily the amount he *has* to pay; that is, these vertical distances we have been describing are not insurance *prices*, which are determined by the supply conditions in the risk acceptance

industry as well as consumer preferences for avoiding risk. It is reasonable to interpret Hesiod's exhortation to the Boeotian farmer to "Take fair measure from your neighbor and pay him back fairly with the same measure, or better" (*Works and Days* 349–351) as a recommendation that farmers implicitly insure one another – the "or better" being a risk premium to increase the chances of getting a helping hand the next time an opportunity arises.

3.14 Irrational Behavior

It apparently is easy to toss out the vague objection, "Yes, but you've assumed rationality." We introduced this chapter with the elements of the definition of rationality used in contemporary economics. That may or may not satisfy individuals determined to find simple reasons for not embarking on a study of contemporary economics to fortify their studies of the economic behavior of people in ancient Mediterranean and Aegean societies. While economics offers specific definitions of what it means by "rational" behavior, it has been within the spirit of discussion in contemporary economics to offer specific examples of what might constitute "irrational" behavior and study its consequences.

It is simple to show graphically that three important categories of nonrational, individual behavior still yield downward-sloping aggregate ("market")

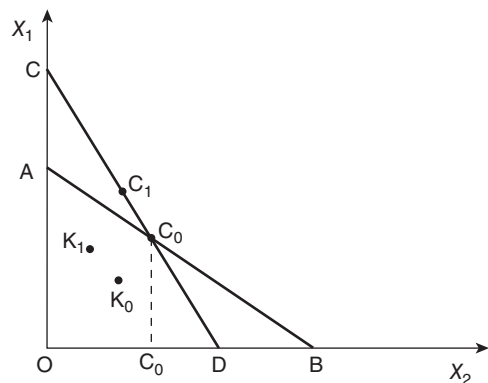


Figure 3.29 Irrational behavior: impulsiveness and inertia.

demand curves.¹³ Impulsiveness, inertia, and failure to maximize can be represented in Figure 3.29. Impulsive behavior is characterized by random choices of goods X_1 and X_2 ; choices might be determined by throws of a die, but constrained in Figure 3.29 to the consumer's opportunities, which are defined by the budget lines AB and CD in alternative situations. If we initially restrict all our consumers to reach the budget line AB (they are still maximizers, just random ones), the average, or "expected" combination of X_1 and X_2 (the centroid of individual consumption choices) will be at the center, c_0 , of the initial budget line, AB. Rotation of the budget line through the initial consumption point is equivalent to changing the relative prices of X_1 and X_2 but keeping income constant; budget line CD represents this increase in the relative price of X_2 . Again, the expected (average or centroid) consumption will lie at the center of the new budget line, point c_1 , which is a move toward the good whose price has fallen. Area c_0DB is no longer available to consumers, so they are pushed up into the new area which now is accessible to them, c_0AC . We could continue to change relative prices, and the centroids of random consumption would trace out a negatively sloped demand curve.

It is common to appeal to tradition and custom when explaining the consumption choices of noncontemporary non-Westerners, so the case of inertial consumers has considerable relevance. In this model, once consumers make a choice of X_1 and X_2 , they do not change that choice as long as the portion of the opportunity set (the budget) in which their consumption lay remains available. Begin with the maximizers so that all the consumption choices lie along the original budget line, AB. Suppose that the average consumption of many consumers is again represented by point c_0 , although we do not deduce that point from an averaging. Once again, change the relative prices to budget line CD. The consumers who originally chose X_1 , X_2 combinations on the stretch of the budget line between A and c_0 remain there, but those in region c_0DB can no longer attain their original consumption. All of those who are forced to change their consumption choices had consumption of X_2 above the average of Oc_0 . Some of these consumers would have their choices forced

back at least to the stretch of the new budget line between c_0 and D; some might make choices in the newly opened region c_0AC , but if none of them did, the new average consumption of the entire group of consumers ("the market") would fall somewhere on c_0D , and the aggregate demand curve still would slope downward.

Consider inefficient consumers – those who leave possible consumption "on the table," so to speak. Take the random choosers first. With the initial prices, the centroid of consumption for them would be at the centroid of the entire opportunity set OAB, point k_0 . At the changed set of prices, their consumption centroid would be the centroid of opportunity set OCD, which, by the geometry, must lie to the "northwest" of k_0 , at k_1 . A continual shifting of the relative prices would trace out a downward-sloping market demand curve. Inefficient, inert consumers would be distributed throughout the interior of the opportunity set AOB but those choosing in area c_0DB would find themselves pushed out of those consumption choices back behind the portion of the line from c_0 to D. Again, the market demand curves comprising the choices of all the individuals choosing these two goods slope downwards.

The significant point to be taken from these simple demonstrations is that the decision rules that individuals use have at best a second-order influence on the downward slope of the demand curve. It has been the changes in the opportunity sets that have produced the downward-sloping demand curves. If we call these "market averages" of consumption choices of "representative" consumers, then the representative consumer behaves rationally, satisfying the components of rationality described in the opening section of this chapter. Comparable reasoning shows that firms also need not be active maximizers to have their demands for inputs slope downward in their prices. Once they depart from profit maximizing behavior, they effectively are constrained by a budget just as consumers are. If they continue to exceed that budget constraint they eventually will use up all the resources at their disposal. This characterization applies to governments and government agencies or firms as well as private producers.

3.15 Fixed Prices

We often hear it said and see it written that prices in antiquity were fixed – they never moved, for decades at a time. Karl Polanyi made the contention in 1944 for the Mesopotamian region, and J.J. Janssen's study of the price records among the artifacts from Deir el-Medina found that a number of specific prices changed little and seldom over a period of eighteen to twenty decades.¹⁴ Economists have considerable difficulty believing such a contention, even when confronted with supporting evidence such as that offered by Janssen. Such an attitude may strike the ancient historian, archaeologist, and philologist as dogmatic and worthy of their dismissal. Perhaps it would be useful for the dialogue to explain the economic reasoning behind such stubbornness.

Figure 3.30 shows how an economist would think about the matter of fixed prices. Suppose that in some initial period, supply S_0 and demand D_0 of some agricultural product were equilibrated – that is, the quantity of lentils, say, supplied was equal to the quantity demanded. At this equilibrium, the cost of producing the last unit of lentils (possibly the last hectare's output) – and consequently the cost of producing all of the units when you can't tell which unit is the last and which is the first – is equal to the value that consumers – many of them the producers themselves and their families, but some of

them workers in town – place on that last unit of consumption. This unit production cost and the unit valuation, which we can call a price to save on words, is $\bar{p}$ in Figure 3.30. The quantity both supplied and demanded is q_0 . Now, suppose that there is a particularly dry season and the output of lentils is lower than previously. The supply curve shifts up to S_1 ; that is, the cost of supplying each unit of lentils is higher than it was in the previous season because of the scarcity of water. Let's suppose that the demand for lentils remains unchanged. (Actually, we could make a case for demand falling off a bit if a sizeable portion of the income supporting the demand for lentils came from their production, but we ignore this for the time being.) This means that the demand curve is constant at D_0 .

At this point, we need to digress a moment to discuss the meaning of the supply curve in this case. So far, we haven't explicitly considered the time required to produce something. With agricultural products, farmers initiate the season by preparing a certain amount of ground and sowing so many seeds, on the expectation of what the weather will be between then and harvest time. Their predictive abilities probably aren't bad, especially if, as a society, they've lived and farmed in the area for a number of generations. But their expectations will be correct only on average, which means that each year the weather probably will be at least a bit different from the farmers' expectations. Once in a while – more commonly in particularly dry areas – their expectation will be considerably off. The expectation regarding the weather will translate into an expectation about the price that the crop will command at harvest time, given the farmers' knowledge about their production technology (constant over the period of the growing season) and the demand for the crop (little reason for it to be a whole lot different than last year). Thus, in Figure 3.30, we could say that the farmers growing lentils start season 1 with the expectation that price $\bar{p}$ will prevail at harvest time and they consequently plan to supply lentils according to supply curve S_0 . The weather is worse than expected, and by harvest time, to deliver any particular quantity of lentils to demanders would be more costly than it was last year. It is in this sense that the supply

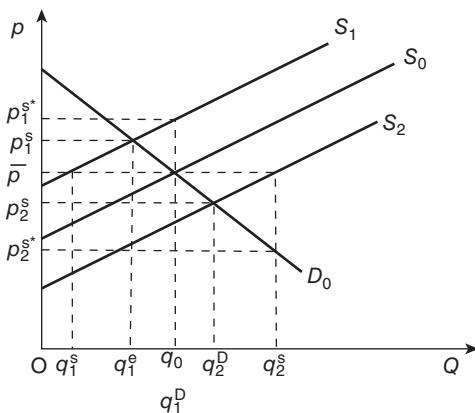


Figure 3.30 Consumption behavior with fixed prices.

curve shifts up to S_1 . Now that S_1 characterizes supply conditions, with the price still at $\bar{p}$, the farmers would want to supply only quantity q_1^S , which is substantially lower than the quantity demanded at price $\bar{p}$. To get them to supply the same quantity as last season would require a price of $p_1^{S^*}$. The consumers' valuation of $p_1^{S^*}$ would elicit a quantity supplied of q_1^E , which would equate the quantity supplied and the quantity demanded, but with price fixed at $\bar{p}$, there is a supply shortfall of $q_1^E - q_1^S$. The fixed price removes any mechanism within the conception of the problem (= "the model") capable of resolving this shortfall. Before we explore ways of resolving this supply shortfall, let's explore the opposite case, of a particularly good crop year.

The supply curve shifts out (or down, if you prefer) to S_2 , and once again, there is no reason for the demand curve to shift – the weather has affected only production, not preferences or income (with the same caveat about income as we made above). This season, farmers would be willing to supply quantity q_2^S at price $\bar{p}$. With their demands unchanged, consumers would be willing to consume quantity q_2^D when this much is available, leaving an excess of $q_2^S - q_2^D$ of production over what people are willing to consume at price $\bar{p}$. If quantity q_2^S were actually made available, consumers would be willing to pay only $p_2^{S^*}$.

Conceivably, with supply conditions represented by S_2 , farmers could leave the difference between q_0 , the amount that consumers would be willing to take at the fixed price $\bar{p}$, and q_2^S , the amount farmers are willing to supply at that price, lying in the fields for compost. But with supply conditions represented by S_1 , we have an immediate problem: "How do the consumers get fed at price $\bar{p}$?" Exactly *who* gets the less-than-demanded quantity q_1^S ? Do some consumers get their ordinary purchase amount and everybody else nothing, or does everyone get his or her ordinary amount reduced by the same percentage? Either solution poses critical difficulties for many of the consumers. If the good we were discussing were one that consumers could wait for, say shoes, this supply shortfall is an ideal situation to be resolved by the formation of a queue: consumers sign up and wait, just like the

Russians waiting for the plumber.¹⁵ Queues for food let consumers spend their valuable time as a substitute for cash: if the price is too low to clear the market (that is, to equilibrate demand and supply), the cash plus the waiting time can be made to equal the quantity of cash that would have cleared the market, but possibly also eroding some consumer surplus, which a flexible price would not have done. Thus, just looking at the posted price does not reveal the full price consumers pay in terms of cash payments plus waiting time.

Of course, this is the point at which the role of stockpiled lentils and Polanyi's redistribution come into the story, but from outside the model and raising the questions of where the stockpile came from and where the allocation rules guiding the redistribution from stocks came from. If prices were ordinarily fixed, bore no relationship to quantities available and quantities consumed, and were not used in redistribution plans, one wonders why prices were noted and written down in the first place. (Remember that Pharaoh *sold* the grain from the storehouses that Joseph had him fill in anticipation of the Biblical seven years of famine in Egypt.¹⁶) Suppose that stockpiles of lentils were available from previous years' excess supplies (at fixed price $\bar{p}$). Ignore for the moment the length of time lentils could have been stored before they deteriorated beyond a consumable condition. Ignore also the mechanisms required to get farmers to transfer excess product to whoever owned the stockpiles (a combination of payment and compulsion, and dependent of course on land and crop ownership and contracts specifying rights to output). Suppose the stockpile owner – the government ("state") – distributes food in the total amount q_0 minus q_1^S equally among the consumers. They might not all have the same preferences, leaving some with a larger endowment of lentils than they want and others with less. As we will see in Chapter 5, this is an ideal situation for a set of exchanges that would bring aggregate demand into equality with aggregate supply. If consumers could put their excess endowments of lentils up for sale in some centralized facility in which all consumers could express their demands (a lengthy way of avoiding using the term "market"), the additional supplies

from stores would bring the total quantity available up to what was consistent with the fixed price $\bar{p}$, and that price should emerge, although the payments would be among various holders of lentils disbursed from stores. (The “price,” of course, might turn out to be combinations of other agricultural products, labor time, pots, shoes and clothing, and so on, depending on what items various consumers might have to hand in the absence of a uniform currency.) Without such a central clearinghouse for excess endowments, pairs of individuals would have to find each other, and the prices at which endowments change hands would vary across pairs of individuals. Possibly, if resales were an option, individuals who acquired their extra lentils at a lower price than they notice elsewhere might be willing to part with some of what they got at a higher price from someone else who still hasn’t found the people with the lower-than-average preferences; eventually, the prices might settle down to something close to $\bar{p}$. These are all “ifs.” Food stockpiles are costly to acquire and maintain. Disbursal rules frequently operate squeakily rather than smoothly, at least partly because the stockpile managers can’t observe supply-quantity shortfalls directly or clearly and consequently don’t know when they are really justified in making disbursals, and to whom. The idea of smooth disbursals from centralized stores, according to need, strikes an economist accustomed to studying their operation, and the incentives involved in their operation, as appealing to a one-for-and-all-for-one mentality that is unlikely to have been sustainable among the ordinary stresses and strains of life in a society of people. We will get some further idea of the information costs of centralized decision making in Chapter 6.

The extensiveness of fixed prices across time and locations, and their alleged prevalence across goods with all sorts of supply conditions (some affected by the weather and some not, some affected by arrivals of foreign supplies, some by competing demands for labor at different times), seem to leave the stockpile explanation of fixed prices – for a hundred or so years at a time – leaned on rather too heavily for credibility. Too many stockpiles! This skepticism regarding stockpiles as a sufficient mechanism for adjusting random variations in production to demands

may not be highly persuasive to scholars heavily committed to belief in fixed prices. The fundamental objection of the economist to a belief in fixed prices is that the existence of stockpiles does nothing to alter the existence or structure of individual preferences (utility, demand, and supply functions), which would continue to operate in the presence of fixed or flexible prices. Whether stockpiles or some other allocation mechanism were used to adjust supply–demand discrepancies, the same forces of excess or dearth would continue to emanate from individual consumers and suppliers.

Appeal to stockpiles and redistribution raises literally more questions than it resolves about the possible fixity of typical ancient prices. Stockpiles – or inventories, which is what they are – can smooth price fluctuations in the face of both supply and demand fluctuations, but their operation is commonly implemented with pricing, as in the case of Pharaoh’s disbursals from his grain storehouses, cited above. It would be possible to substitute another set of rules to govern disbursals from inventories for a rule that said, essentially, “If the price they offer us is high enough to meet or exceed our minimum price [commonly known as a “strike price” in contemporary public inventory holdings], sell to ‘em.” What might some of these other rules contain? One would be “Release quantities sufficient to keep the price constant.” Implementing this would be difficult – the bursar could overshoot or undershoot the price target. Another might be, “Release quantities sufficient to keep all recipients’ consumption levels constant.” Without knowledge of how much recipients already had, this would be impossible to implement. Additionally, in a particularly bad year, it might not be possible. It is reasonable to question whether rules like these could account for an alleged constancy of agricultural prices (not to mention other goods’ prices) over a score of decades, or even a few decades – or even a few seasons. Large inventories would be necessary to keep prices constant in the face of large supply fluctuations – particularly bad or good crop years. The time necessary to build such inventory levels might exceed the length of time that ancient storage technologies could keep agricultural stores edible. If technology did permit such relatively large inventory holdings, the cost would have been quite high.

Think about the cost this way: in a society that derived about 85 or 90% of its total income from agriculture – to be generous, say 75% from storable agricultural products like grains, nuts, and dryable fruits – the cost of maintaining inventories large enough to compensate fully for a 50% crop shortfall would be equivalent to hoarding 37.5% of national income practically in their mattresses. It would take a rich society to be able to afford that. And the inventories would have to be turned over frequently even if they weren't needed to cover crop shortfalls; unless perfectly good food were simply destroyed, the effects of such routine disbursements to make room for fresher stores would affect prices anyway.

3.16 Applying Demand Concepts: Relationships between Housing Consumption, Housing Prices, and Incomes in Pompeii

Andrew Wallace-Hadrill (1994, 98–103, Figures 5.3–5.5) has tried estimating the residential densities of 13 categories of house size in Pompeii.¹⁷ While well aware that the distribution of people across various house sizes was unlikely to have been uniform from the smallest and meanest houses to the largest and most lavish, in the absence of further information, he distributed an estimated population of 10 000 people across the size categories under an assumption of a constant population density over all house sizes and, alternatively, assuming one person per room in houses of all sizes. Despite the shortcomings of Wallace-Hadrill's two procedures, there is insufficient information with which to do any better in estimating either residential population densities or the distribution of personal incomes across the residents of those houses. If floor space consumption information were available, income could be inferred with a housing demand equation (although the elasticity values would have to be decided upon a judgmental, rather than a directly empirical, basis). Alternatively, if income were available for the residents of at least some houses, their floor-space consumption could be projected, contingent on the reasonability of the demand elasticities used. As the data exist, they simply are not sufficient to yield information on residential population density.

We can, however, use the data assembled by Wallace-Hadrill to estimate the spread between the lower and upper ends of the personal income distribution in Pompeii, although we cannot estimate the percentages of the population receiving incomes in between these two extremes. That is, we can't reconstruct the actual income distribution, but there is something of potential interest that we may be able to back out of the information available.

We know from demand theory that higher incomes generally would lead people to consume more housing. While floor space is only one characteristic of a house, the hedonic model of differentiated products gives us a basis for using floor space as primary indicator of the quantity of house consumed. It is reasonable to believe that people with higher incomes would have consumed more floor space per person, and the income elasticity of demand captures this aspect of consumption. As a practical matter, it also seems reasonable to believe that the floor space would have become more costly as we move up the size categories, reflecting, for example, the installation of mosaics, the use of higher quality stone and other building materials, and the generally higher quality construction methods in the homes of people with the income to consume larger floor space (larger houses). We can account for this trend in the price per unit of floor space with the price elasticity, which retards consumption as price increases.

To combine these two aspects of the demand for housing, we give ourselves a floor space demand equation that is a function of income (y) and price (p): $s^d = y^\eta p^\epsilon$, where η is the income elasticity of demand for floor space and ϵ is its price elasticity of demand. For values of η and ϵ we rely on contemporary estimates of the income and price elasticities of demand for housing in the United States: $\eta = 0.70$ and $\epsilon = -0.75$ (see the discussion of housing demand in Chapter 12).¹⁸ Of course there is no illusion that these are the "true" values of these parameters for ancient Pompeii, but their relative magnitudes accord with consumption theory and their absolute magnitudes are reasonable *ipso facto* and accord with contemporary evidence from around the world.

We don't know how many people consumed how much floor space, but it is safe to presume that *someone* consumed what Wallace-Hadrill

characterized as the average floor space of the entire city if the population at the time of destruction was 10 000, about 35 m². The smallest of Wallace-Hadrill's 13 size categories is 0–99 m², so this is somewhat below the mean of that category. Clearly this would characterize the floor space consumption of a family close to the lower end of the personal income distribution. To solve for the income, in terms of floor space valued at the quality used in the size-category-1 houses, we set 35 m² = $y_1^{0.7} p^{-0.75}$. We can set $p = 1$ to represent the price of the lowest quality housing. Solving for y yields 160.6: the lowest income family has an annual income that would rent them 160.6 m² of floor space, although they choose to rent only 35 m².

To estimate the income of a family at the opposite end of the income spectrum, we allow for change in the unit price of floor space between houses in size categories 1 and 13 by supposing that the unit price of floor space increases by 10% and, alternatively, by 5% between each building size category. We also compare the consequences of a less price-elastic demand for housing at the high end of the income scale: $\epsilon = -0.20$ as well as $\epsilon = -0.75$. We consider the family consuming 2500 m² of floor space, which is the midpoint of size category 13. For a price increase of 10% between size categories, the magnitude of the price variable in category 13 will be $1.1^{12} \approx 3.14$. A 5% price increase between each size category yields a price in category 13 of $p \approx 1.79$. Table 3.1 shows the income calculations for the wealthy household under the four possible combinations of relative price increase and price elasticity of demand.

The reasons these parameter differences make so much difference in the income estimates are instructive. Consider the 10% cross-size category price increase. With the more elastic price elasticity, the price increase would reduce the demand

for floor space by more than would occur with the less elastic value. Remember that the quantity of floor space consumed is given at 2500 m². A larger income would have been required in combination with the larger (in absolute value) elasticity to leave the family consuming 2500 m² of floor space, when the unit price of floor space is three times the lowest quality level. Notice that if the materials in the category-13 house are less costly ($\Delta p = 0.05$), the income required to consume that size house is reduced considerably when the price elasticity is -0.75 , and proportionally by not nearly as much when it is only -0.20 . More generally, these results show that if price elasticities of demand for housing were quite low among the wealthy Pompeians, the richer houses we observe in the remains would imply families of lower income than if we believe we are observing the actions of people who were relatively more cost-conscious.

The spreads implied between the incomes of the very poor and the very wealthy in Pompeii are quite wide: by a factor of 1516 times in the case of the high price increase and the larger price elasticity, or still by a factor of 526 times in the case of the smallest spread, with the low price increase and the smaller price elasticity. Are these spreads too wide to be believable? A quick comparison with contemporary American income differences may help answer that question. Consider a relatively poor American making \$10 000 per year; the lowest spread would have an American at the top end of the income distribution (say, in a town comparable to Santa Barbara, California?) making \$5.26 million per year, say, a reasonably but not extraordinarily successful lawyer – probably not a “one-percenter”. The highest spread would have the family at the top end making \$15.16 million per year, compared to the poorer countryman making \$10 thousand. These spreads are within the contemporary experience and, we believe, probably could have been well within the ancient Pompeian experience.

Now for two elements of inexactitude in these calculations. First, as Wallace-Hadrill pointed out in considerable detail, the larger houses – the residences of the wealthier families – frequently contained sleeping quarters for variable numbers of retainers and slaves. He was unable to control for this factor, and we can do no better.

Table 3.1 Wealthy income in Pompeii for alternative parameter values.

	$\Delta p = 0.10$	$\Delta p = 0.05$
$\epsilon = -0.75$	243 437	133 853
$\epsilon = -0.20$	99 109	84 498

However, the ownership of the floor space calculated with the demand equations resides with the wealthy family and we could interpret the family's demand for space to include the quarters for the retainers who provided services around the residence. Surely, the residential square footage of the vast majority of the retainers would have been considerably smaller than that of their wealthy employers/masters. Second, some of the floor space in Wallace-Hadrill's original estimates may have been work space rather than what we would consider living space. If wealthy people's derived demand for work space¹⁹ were characterized by about the same parameter values as their demand for living space, there would be little error on this account.

The foregoing thoughts have not distinguished between the demand for purchased housing and the demand for rental housing. If a consumer purchases housing, the dwelling provides not only a flow of housing services but a capital asset that may (or may not) increase in value, adding a portfolio component to the demand to own housing stock. Both markets clearly existed in antiquity although the evidence is more ample for Rome than for Classical Greece. Cahill (2002) did not report any evidence regarding rentals versus ownership at the short-lived, Chalchidian city of Olynthus, but evidence from ancient Italy is ample. Renters in Rome during the early Imperial Period included upper class families as well as people barely making ends meet, with the sorting (at least in second-century C.E. Ostia) being vertical within buildings rather than horizontal as in most Western cities today (Frier 1980, 14–16, 39–47). Developing a set of rules of thumb for distinguishing rentals from owner occupancies in two insulae in Pompeii and Herculaneum, the cities covered by the eruption of Vesuvius in 79 C.E., Pirson (1997, 173, 175–177, 181) identified some 100 rented *tabernae* (workshops, with or without living space) and *cenacula* (upstairs rental flats) in Pompeii.²⁰ Pirson, using his criteria for distinguishing rentals from owner-occupied dwellings, estimated that 16% of units in Pompeii were *cenaculae*, and 26% in Herculaneum, at the time of their burial (Pirson 1999, 9). Adding in the *tabernae*, Pirson estimated 42% of units in Pompeii to have been rentals and 53% in Herculaneum (Pirson 1999, 175). DeLaine (2004) uses

close reasoning from a combination of Hillier and Hanson (1984) and Hanson (1998) spatial syntax measurements, more conventional architectural characteristics (absolute and relative sizes of units and rooms), decorative / symbolic features and historical information to form reasoned conjectures regarding economic, social and family status of original and subsequent occupants of medianum apartments in the southeast corner of excavated Ostia, including assignment of units to owner-occupation or rental. The spatial syntax measurements develop hedonic characteristics more subtle than room number, type, and size, extending to methods in which spaces and connections conferred such valued attributes as privacy and control of movement within a dwelling unit. Stöger (2011, 162, 171) has used these measurements in an insula at Ostia to construct indicators of neighborhood quality, another hedonic characteristic that would attach to individual dwelling units. Laurence (1994, Chapter 7) uses space syntax calculations to study property types and interactions in two “structurally different” (p. 121) parts of the city, although his results remain difficult to interpret.²¹

Taking a small leap from the occupation of houses to the uses and activities conducted in them helps tie together a number of economic concepts brought up so far: consumption, production, household production, variety. Allison (1999) has called into question a number of common practices in the archaeological analysis of Roman domestic artifacts in a manner that highlights these interrelated economic issues. First, without detailing her concerns or her procedures of addressing them, she encourages closer relations between the production and consumption of these artifacts than generally has been achieved to date. I doubt that thinking jointly about the supply (production) of and demand (consumption) for artifacts will yield further information about prices or costs (or probably only quite roughly), but thinking about the determinants of supply and consumption when analyzing the use of ancient domestic artifacts cannot but increase the insights brought to bear on them. Second, Allison's emphasis on the analysis of entire assemblages, with particular attention paid to their contexts, as contrasted with catalogs of similar artifacts from an excavation – or series of

excavations – is the archaeological equivalent of thinking in terms of household production, and can even extend to the economic allocation of time. And third, her attention to the possibilities of changes in uses of similar types of artifacts over lengthy periods of time, as well as the details about use of individual artifacts contained in various aspects of their remains (size and shape as well as such “acquired” characteristics as discolorations) may be thought of as issues in the economics of variety. Altogether, her analysis is a useful complement to the application of spatial syntax analysis of buildings and neighborhoods by DeLaine, Lawrence, and Stöger.

3.17 The Economics of Mycenaean Vases, II: Demand

In our previous discussion taking Wijngaarden’s topic of Mycenaean vases in Ugarit as the point of departure, we talked somewhat “at the other end” from those four authors’ principal interest, the value of the three vase types – small stirrup jars, amphoroid kraters, and conical rhyta. In this chapter, it’s appropriate to deal directly with the same issues, though we’ll go at them a bit differently. Wijngaarden’s working definition of *value* was, roughly, the desirability of an object, modified by its accessibility, which makes value, as Wijngaarden notes, as much a characteristic of people viewing items as of the item itself, if not more.²² This concept of value corresponds sufficiently with the concept of it we’ve used in this chapter that we will be talking about the same characteristic of these vases in the following discussion.

Wijngaarden’s approach to investigating the value of these vases is to examine the contexts in which they were found in hopes of finding some indirect evidence of valuation. Of course, this is a standard analytical technique in archaeology and has considerable merit in using information on related objects to infer additional information about some object of particular attention. Remember that demand for any good is a function of prices and income – the price of the good in question and the prices of substitutes and complements. As he proceeds, most (but

not all) of Wijngaarden’s inferences from context refer to the incomes of apparent users of these vases. In the cases of the conical rhyta, however, Wijngaarden notes several instances of the Mycenaean rhyta being found together with locally produced clay rhyta, and in the case of one in the House of the Hurrian Priest together with “a stone libation tube, which could fulfill a similar function” (21); in his commentary, De Mita suggested the substitution of these ceramic rhyta for metal ones (25). All three instances involve substitutes for Mycenaean rhyta: local ceramic rhyta, presumably a little cheaper, if only because of a smaller transportation cost component; stone implements that could accomplish the same ritual tasks, presumably somewhat more expensive because of the time involved in working their material, although they probably included less transportation cost; and the metal rhyta which presumably were considerably more expensive because of both their material and the skills involved (presuming, possibly unreasonably, that metalworking skills were more costly to develop – and hence cost more to employ; see Chapter 10 on labor – than potting skills). Chances are, the “value” of the Mycenaean rhyta was bracketed by the “values” of these substitutes. You’ll note, of course, that we have slipped in here cost-based estimates of what values would have been. What makes this legitimate?²³ Remember also that the revealed value of an item is the price that emerges from the intersection of a supply curve and a demand curve. We’ve been moving up and down a supply curve, implicitly suspecting that somewhere within the range of costs (prices) we’ve introduced with these substitutes, the demand curve for the Mycenaean rhyta must have crossed it.

Let’s move a step back in our analysis at this point, to ask how we would measure the value of these Mycenaean vases. In their discussion of what conferred value, Wijngaarden, and Whitelaw in his commentary (34), moved toward the relatively philosophical concept of “meaning” of an object, and away from a practical, workaday sort of definition of value as something revealed literally on a daily basis by the actions of consumers. No numeraire was offered in terms of which to measure “meaning,” or

compare that quality or characteristic of one object relative to that of another object. In this chapter, we've stressed that money is unnecessary as a numeraire – that a local, Ugaritian conical rhyton will do just as well (or almost as well; the local rhyta might vary in their size and quality, leaving us with the problem of finding a particular one to refer to so that all our references meant the same thing). What would comprise a practical numeraire, or set of numeraires, for our endeavor to assess the value of Mycenaean vases in Ugarit? Shekels are probably out. There's no unique, correct answer to this question; only more or less useful answers. It's possible that the best numeraires, which will remain imprecisely defined, will be the array of substitutes for these vases. For the conical rhyta, we've mentioned three substitutes; probably the ceramic ones would be the most practical numeraire – or cost / value reference – because of the similarity of material. The amphoroid kraters tended to be found in apparent kitchen contexts, sometimes with storage vessels. Those local vessels that would (or could) have performed the same services would appear to be the most satisfactory numeraires. The differential artistry of decoration of the Mycenaean and local vessels would have to be considered, but the indiscriminate mixture of the finely decorated Mycenaean examples with plain, local storage vessels would not yield a high estimate of the differential valuation of the former's decoration. That is, the hedonic price of the decoration would not seem to have been high, based on the contextual associations (another example of a price inference from context). However, we must remember the potentially lengthy use of these kraters, which makes them examples of durables rather than simple "consumer" goods. The Mycenaean kraters could have started their "careers" at Ugarit in one use and have been switched to another use before they wore out (were broken), possibly because the original valuation of the decoration had deteriorated – or possibly because the decoration itself had deteriorated, making the vase, in the eyes of a person with taste in that sort of thing, a "formerly decorated" vase that would still hold things that needed holding.

The concept that "power" was the entity that "restricted" access to some of these Mycenaean vases can be simplified with the concept that purchasing power, or income, would have been sufficient to ration these vases to the people or families in whose contexts they were found. Similarly, the idea that alternative value systems operating in Ugarit at the same time would have implied different values for different people can be interpreted somewhat differently with the use of an aggregate demand curve for any of these vase types. Remember that we add individual demand curves horizontally, which yields the result that at any particular price, the total quantity demanded of an object is the sum of what each individual would want to purchase at that price. Once we've added the individual demand curves horizontally, we can drop the supply curve across the aggregate demand curve to get a single intersection that defines the price that all individuals in the market (the market is Ugarit in this case) would have to pay, together with the sum total of them that the different individuals would purchase at that price. Thus, even though the residents of Ugarit, taken as a group, might have had radically different "value systems," when it came to buying Mycenaean pots (or any other pots for that matter), they would have competed with one another, with the result that they all would have had to pay the same price for a given type (and quality) of vase. Those who valued them more would have bought a larger number of them; those valuing them less would have purchased fewer; those valuing them still less, but still "valuing" them, could have purchased none. It might be objected that "power" permitted some people to deny access to these vases through some administrative means, such as import restrictions or restrictions on who was allowed to purchase the vases (Wijnngaarden, 4, 22). The former restriction simply would have driven up the price of Mycenaean vases for everyone, while the latter action would (or could) have created a black market, in which official and unofficial prices coexist, with the goods generally being unavailable at the lower, official price.

Using close substitutes as numeraires for each type of Mycenaean vase can help us assess

roughly the relative valuation of the Mycenaean vases to the numeraires, but pretty much leaves us with little to go on as far as comparing the valuation of one Mycenaean vase type to another Mycenaean vase type, much less to other categories of goods altogether. How could we bridge this gap? If we could construct comparative assessments of the costs of different vase types, such as we did with the cost function in section 2.11, how could we be assured that these costs (what people would have *had* to pay, at rock bottom, or else the producer would have sold the good at a loss) would correspond to the prices that people would have been *willing* to pay? It seems obvious that just because you might be able to get away with paying one price for some item it doesn't necessarily mean that you wouldn't have been willing to pay a lot more. There are, indeed, justifications for both the latter suspicion (the concept of consumer's surplus, for one thing), and the former desire to drive price down to cost, at least under a wide range of conditions, but explaining that will take us into the subject of competition, which remains to be introduced in Chapter 4. Suffice it to say that, under the technological circumstances of pottery production, it is probably pretty safe to make the equation between cost and price. Equipped with such an ability to equate production costs with price, we could develop a system of chained comparisons, moving from one type of commodity to another, based on assessments of relative production costs (when actual production costs are lost forever).

Finally we come to the perspective of demand for variety. Why would a resident of Ugarit have purchased a Mycenaean small stirrup jar when he could have had all the small oil containers he wanted from local sources? For one thing, they were aesthetically attractive, and they looked different from the local substitutes for another. Two reasons for buying them, according to the demand for variety. Further "meaning" may have existed, but it is unnecessary to account for the presence of the stirrup jars in even relatively humble domestic contexts, in which many of them were found. Even if they cost a bit more than local substitutes, some of them still could have been purchased.

We've referred to substitutes for the Mycenaean vases several times in this discussion. The substitute may seem like a wildcard – something to be played whenever all other explanatory devices have failed (alternatively, the last refuge of a scoundrel). However, recall from section 3.5 that the substitute concept can be embodied in an elasticity, and that demand theory puts clear restrictions on the values that own-price, cross-price (the relationships between the Mycenaean vases and their substitutes and complements), and income elasticities of demand can take, given expenditure shares. Of course we will never locate data to tell us what those elasticities were, but we could explore a bit numerically, an exercise to which elasticities lend themselves well by virtue of the restrictions on their magnitudes and the interpretability of numerical values above and below certain points (zero and plus or minus one are key values; see section 3.5 again if you don't remember why). We could hypothesize some own-price and income elasticities for both Mycenaean vases and local substitutes; then suggest some budget shares we think to have been reasonable. Put those together in the proper formula and solve for the cross-price elasticity, which will be the only parameter left to vary within the constraint imposed by the elasticity relationship.

So, after all this discussion, we still can't say exactly how much the residents of Ugarit valued Mycenaean vases, in terms of either shekels or any of our more plausible numeraires, but we can make some sense of some of the patterns observable in the archaeological data without appeal to complicated "social strategies" or the indefinable "meaning" people attach to the objects in their lives. The prices of some of these vases could have restricted most of their purchase and subsequent ownership to people with higher incomes, without recourse to appeal to social restrictions, which may have been largely unenforceable anyway. Some people ("all" are not necessary) probably would have been willing to pay a bit more for an attractive additional variety of an object of which they already possessed five. People's everyday acts would have been sufficient to reveal their values regarding these

objects, if not necessarily their assessments of the meaning of life in general. And when we want

to talk about values, we need something for a measuring rod.

References

- Allison, Penelope M. 1999. "Labels for Ladles: Interpreting the Material Culture of Roman Households." In *The Archaeology of Household Activities*, edited by Penelope M. Allison. London: Routledge, pp. 57–77.
- Allison, Penelope M. 2001. "Using the Material and Written Sources: Turn of the Millennium Approaches to Roman Domestic Space." *American Journal of Archaeology* 105: 181–208.
- Becker, Gary S. 1962. "Irrational Behavior and Economic Theory." *Journal of Political Economy* 70: 1–13.
- Becker, Gary S. 1965. "A Theory of the Allocation of Time." *Economic Journal* 75: 493–517.
- Cahill, Nicholas D. 2002. *Household and City Organization at Olynthus*. New Haven CT: Yale University Press.
- DeLaine, Janet. 2004. "Designing for a Market: 'Medi-*anum*' Apartments at Ostia." *Journal of Roman Archaeology* 17: 147–176.
- Dixit, Avinash K., and Joseph E. Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67: 297–308.
- Frier, Bruce W. 1980. *Landlords and Tenants in Imperial Rome*. Princeton NJ: Princeton University Press.
- Hanson, Julianne. 1998. *Decoding Homes and Houses*. Cambridge: Cambridge University Press.
- Hillier, Bill, and Julianne Hanson. 1984. *The Social Logic of Space*. Cambridge: Cambridge University Press.
- Houthakker, H.S., and Lester D. Taylor. 1970. *Consumer Demand in the United States: Analyses and Projections*, 2nd edn. Cambridge MA: Harvard University Press.
- Janssen, Jac J. 1975. *Commodity Prices from the Rames-*sid* Period: An Economic Study of the Village of Necropolis Workmen at Thebes*. Leiden: E.J. Brill.
- Keith, Kathryn. 2003. "The Spatial Patterns of Everyday Life in Old Babylonian Neighborhoods." In *The Social Construction of Ancient Cities*, edited by Monica L. Smith. Washington: Smithsonian Books, pp. 56–80.
- Lancaster, Kelvin. 1979. *Variety, Equity, and Efficiency; Product Variety in an Industrial Society*. New York: Columbia University Press.
- Laurence, Ray. 1994. *Roman Pompeii; Space and Society*. London: Routledge.
- Lluch, Constantino, Alan A. Powell, and Ross A. Williams. 1977. *Patterns in Household Demand and Saving*. New York: Oxford University Press.
- Musgrove, Phillip. 1978. *Consumer Behavior in Latin America: Income and Spending of Families in Ten Andean Countries*. Washington, D.C.: Brookings Institution.
- Pirson, Felix. 1997. "Rented Accommodation at Pompeii: The Evidence of the Insula Arriana Pol-*liana* VI 6." In *Domestic Space in the Roman World: Pompeii and Beyond*, edited by Ray Laurence and Andrew Wallace-Hadrill. *Journal of Roman Archaeology* supplement. Portsmouth NH: Journal of Roman Archaeology, pp. 165–181.
- Pirson, Felix. 1999. *Mietwohnungen in Pompeji und Herkulaneum; Untersuchungen zur Architektur, zum Wohnen und zur Sozial- und Wirtschaftsgeschichte der Vesuvstädte*. Munich: Dr. Friedrich Pfeil.
- Polanyi, Karl. 1944. *The Great Transformation*. New York: Rinehart.
- Shepherd, James F., and Gary M. Walton. 1972. *Shipping, Maritime Trade and the Economic Development of Colonial North America*. Cambridge: Cambridge University Press.
- Stigler, George J., and Gary S. Becker. 1977. "De Gustibus Non Est Disputandum." *American Economic Review* 67: 76–90.
- Stöger, Hanna. 2011. *Rethinking Ostia. A Spatial Enquiry into the Urban Society of Rome's Imperial Port-Town*. Leiden: Leiden University Press.
- von Neumann, John, and Oskar Morgenstern. 1947. *Theory of Games and Economic Behavior*, 2nd edn. Princeton NJ: Princeton University Press.
- Wallace-Hadrill, Andrew. 1994. *Houses and Society in Pompeii and Herculaneum*. Princeton NJ: Princeton University Press.

Suggested Readings

- Becker, Gary S. 1971. *Economic Theory*. New York: Knopf. Chapters 1–3.
- Friedman, Milton. 1976. *Price Theory*. Chicago IL: Aldine. Chapter 2.
- Pindyck, Robert S. and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapters 2–4.

Notes

- 1 Experience indicates that many people may have difficulty believing that transitivity of preferences is not commonly, even routinely, violated in practice. If you share this difficulty, try the following exercise. Identify some pair of choices between which your own preferences are clear – where you prefer *A* to *B*. Next, identify some other choice *C*, to which you prefer your previous choice *B*. Now, see if you prefer *C* to *A*. Repeat this exercise until you have identified three such cases in which you prefer *A* to *B* and *B* to *C*, but *C* to *A*. Finally, submit your intransitive preference examples to a friend and ask for a critique.
- 2 In particular, we are ruling out strategic situations with multiple rounds of picking and choosing, in which a bargainer may find it useful to send incorrect signals to a partner or adversary. The situation under consideration is one of a simple, one-time choice.
- 3 A price that appears in a market (or even a non-market situation) need not correctly reveal the true resource costs of the good associated with it. When such a situation occurs, we say that the “market” prices are “distorted.” To give a concrete example, unionization of labor in some activities may raise the wage rate of unionized labor above its opportunity cost. In this case, the unionized wage would be said to be distorted. Taxes also can distort prices. For example, say, a 400% tax on an imported widget which is otherwise a perfect substitute for a locally made widget (the classic “prohibitive” tariff) would distort the information content of the local price of the imported widget (as the tariff indeed is intended to do), implying that the real resource cost of that widget is four times the resource cost of the local variety, when in fact it is exactly equal to the resource cost of the local one.
- 4 Of course, it frequently can fall short of giving unambiguous, qualitative answers. For example, if both supply and demand curves shift in the same direction, the resultant price may be indeterminate although we know that the quantity transacted increases. Correspondingly, in cases in which supply and demand shift in opposite directions, we can predict the movement of price but the new quantity of transactions will be ambiguous.
- 5 A monopsonist is the equivalent to a monopolist, only it exerts its market power in the market for factors of production rather than in its output market. Strictly speaking, a monopsonist would be the only demander of some factor; in practice that would be unusual, but some particularly large producers might still exert some influence over the prices of their inputs. In general, a buyer with monopsony power buys a large enough proportion of the good or factor in question to push up its price by its purchases. Being aware of this influence, it will try to avoid pushing up the prices it faces. A monopsonist need not have such power over all the factors whose services it purchases. We will discuss these phenomena further in the chapter on competition.
- 6 This example is taken from Shepherd and Walton (1972, 21–22).
- 7 This “safeness” also assumes that, if money is being used as the numeraire, either the same money is being used in the locations of comparison or that the proper conversions across currencies have been made. With metal coinage, such conversions are relatively easy.
- 8 Including housing would prove difficult if there were no rental housing. Annual rental prices would considerably simplify the imputation of the user cost of owner-constructed and owner-occupied housing. We discuss the concept of the user cost of a durable good below.
- 9 Actually, a number of statistics are calculated from the components of the error term, *e*, but description of them would take us far from our present purposes.
- 10 Monopolistic competition, which we’ll address in Chapter 4. This is a modification of the “perfect” or “pure” competition among sellers of a product which is essentially identical, regardless who makes it; it allows for differences among products of different producers that consumers can observe and for which they might have a preference vis-à-vis the varieties of other producers. Without saying a lot more about both the perfect competition, monopoly, and monopolistic competition models of how sellers behave, we can’t define much further here.
- 11 Becker won the Nobel Prize in economics for this work in 1992.
- 12 This is a good opportunity to discuss once again, the lack of necessity for these “purchased” inputs to really be purchased, even bartered for, much less purchased at something like a department store for coins or paper money. The same household *using* these cups might also *produce* them. They still will keep a close eye on the marginal benefit of using the cup in whatever consumption activity it plays a supporting role and the marginal cost (possibly in terms of sleeping or fishing) of *making* the cup.
- 13 The original exposition is in Becker (1962).

- 14 Polanyi (1944, 51–52) mixes together “the kingdom of Hammurabi in Babylonia and, in particular, the New Kingdom of Egypt,” focusing on the redistribution and issuance of rations, noting both the absence of money and the use of metal currencies for the payment of taxes and salaries, but without direct reference to either prices or price stability. Janssen (1975, 550–558) summarizes his examination of prices from Deir el-Medina as showing substantial fluctuations in grain prices; “it may possibly be suggested . . . that fluctuation in the prices of grain is reflected in those of oil” (in the Twentieth Dynasty); and “other commodities such as garments, basketry, cattle, and so on, do not show signs anywhere of a regular fluctuation in prices. Higher and lower values occur at random, with no apparent system.” And he notes a constancy in wage payments in grain during the Twentieth Dynasty – despite the record of fluctuations in grain prices. This, of course, is equivalent to a fluctuating *real* wage, even if the quantities of grain paid as wages remained absolutely constant. Pursuing this line of reasoning, as long as the price of grain fluctuated, the *relative* prices of all other products fluctuated correspondingly. Conflating the concept of a level of prices (a topic we have touched on briefly in section 3.8 and which we will examine more closely in the chapter on money and banking, since it is a monetary, as contrasted to a “real” phenomenon) and relative prices, Janssen concludes in the face of the clear evidence to the contrary he has just summarized – and presented in detail in the rest of the book – that prices hardly moved at all – and that when they did, it was not because of what he calls “the influence of rational economic laws” – what we have characterized as supply shifts (weather, cattle trampling crops, illness, and so forth) or demand shifts (sharp changes in income, gradual changes in preferences) – but because of “irrational factors such as the desire for a particular object, skill in haggling, the pride of the maker in his product” – factors for which the economic theory we have related so far routinely accounts, but the factual basis of which he has not reported from his evidence. He notes that the quantity of particular products for which a price is indicated on a papyrus or ostrakon may not be constant, thus accounting for what looks like fluctuating prices, but does not consider the possibility that changing quantity “bundles” could mask price fluctuations equally well; the incentives for making price fluctuations are clearer than for giving the appearance of changes. It is important to emphasize that Janssen’s price observations, like most ancient price observations, are from actual transactions; they are not the ancient equivalent of contemporary “posted” prices, which might be little more than fictions, forcing us to think in terms of shadow prices or the equivalent of modern black-market prices. While the quotations may contain noise – and possibly a serious amount of noise – by virtue of being the products of unit prices and quantities transacted, they are not simply propaganda. And while the small sample of remaining prices may be subject to the occasional “mistaken” transaction – in which one party or the other regrets the terms of trade for one reason or the other – there is no reason why these unit-price “outliers” should be overrepresented in what remains. Altogether, it is not clear that Janssen’s “evidence” of price stability should be characterized as evidence of price stability.
- 15 Mikhail Gorbachev’s joke from 1990.
- 16 Genesis 41:56–57.
- 17 The size categories increase in 100 m² increments, from the smallest size category of 0–99 m², through a tenth category containing houses of 900–999 m²; the next two categories contain ranges of 500 m² each, 1000–1499 m² and 1500–1999 m², and the largest category ranges from 2000–3000 m². For the distribution of houses by size, see Wallace-Hadrill (1994, 77–79), Figures 4.7–4.9 and Table 4.1.
- 18 Expenditure elasticities of demand for housing (equivalent to income elasticities of demand) from several Latin American countries range from 0.75 to 0.92 (Musgrove 1978, 196, Table 6-1). Lluch *et al.* (1977, 545, Table 3.12) estimate total expenditure elasticities (equivalent to income elasticities in the Linear Expenditure System, a method of estimating demands for all goods simultaneously) for 17 developed and developing countries. Countries exhibited considerable variability in their expenditure elasticities of demand for housing, but grouping the countries by per capita GNP in 1970 U.S. dollars to form averages of this elasticity for four income groups yielded an elasticity of 1.01 for the lowest income countries (\$100–500/capita); 0.68 for the next lowest group (\$500–1000); 0.92 for the second highest-income group (\$1000–1500); and 1.24 for the highest (\$1500+); with an overall average of 1.00 for all countries. The definition of housing expenditures includes “housing operation, rent, water, fuel, and light,” not simply explicit or implicit rent (p. 37). The interpretation of these numbers would be that consumers in, say, the second lowest-income group would spend 68% of each additional dollar (denarius) of income on housing.

- 19 Which would have been derived demands, based on the productivity of those structures.
- 20 Allison (2001, 187–188) has criticized Pirson's use of ancient terminology in his sorting procedure but concedes that he may be correct in his identification of spaces referred to in inscriptions as being for rent.
- 21 Keith (2003, 64) uses access graphs to help identify the uses of buildings in a number of Old Babylonian cities.
- 22 Unfortunately, Wallace-Hadrill believes that the correspondence of desirability and access are only superficially analogous to the contemporary economic concepts of supply and demand, but that need not detain us here in applying those concepts directly where Wallace-Hadrill was reluctant (and the discussants applied them only by default).
- 23 A case in which using cost to estimate valuation is not legitimate is the valuation of reductions in pollution, a topic addressed in Chapter 6. The cost of abating the emission of a pollutant is easily observable, and sometimes has been used (incorrectly) as a surrogate for the valuation people exposed to the pollution would place on reducing the emissions, which need bear no relationship to control costs. In fact, some control costs could be so expensive, and the pollutant insufficiently obnoxious, that attempting to use that control technology to abate that emission would be economically nonsensical.

Industry Structure and the Types of Competition

The term *industry* may seem out of place in the study of the ancient world to some scholars, but it needn't bring images of belching smokestacks and speeding bullet trains. In the economic lexicon, an industry is simply a group of producers of the same, similar, or otherwise related, products. It is straightforward to see how producers of exactly the same product – say, sandals – could be considered to comprise an industry, but “similar” needs some further definition. If a producer of one good, using a particular production technology, can switch to production of a related good with little cost of retooling – say, from sandals to harness – the behavior of harness production is likely to parallel the behavior of sandal production, in its responses to technological developments and input and output prices.

One of the principal characteristics of an industry, besides what it produces, is the number of producers. The number of producers of a good is affected by the good's technological characteristics. If there are virtually no fixed costs to producing a good, a producer can supply a pretty small quantity of this item at little or no disadvantage relative to somebody else who supplies a lot of it. On the other hand, if some very expensive equipment is required to produce even the very first unit of a good, a producer will need to be

able to produce quite a few units of it to cover his costs. Consequently, a market – the people who would like to consume this item – of a given size will tend to be supplied by a smaller number of producers who produce at larger scale if fixed costs are substantial.

Producers of identical or close substitute products compete with one another, but how they compete is influenced by the number of producers. In the conditions called perfect competition (section 4.1), lots of producers offer virtually the same product, and competition is very impersonal as a consequence. This condition characterizes quite a few products, ranging from agricultural commodities to many relatively homogeneous manufactured products. The structure of an industry, characterized by the number of producers, influences the relationship between selling price and producer's cost (section 4.2), which is one of the more important reasons to study industry structure.

Monopoly is in the far opposite corner, so to speak, from perfect competition (section 4.3). Something about the good or service being produced lets only one producer survive the competition from other would-be producers. This competition takes the form of adjusting one's prices to run competitors out of business, and when that goal has been accomplished to

the benefit of the surviving producer, sale price moves above the producer's cost to yield a true profit. Not that many goods or services are characterized by genuine monopoly, in the sense of being supplied by only one producer – although scholars have appealed to the monopoly model in discussions of ancient religion, particularly in Bronze Age Crete – and a more realistic characterization of industry structure for some technologically intricate products is oligopoly (section 4.4), in which production conditions allow several producers to survive one another's competition, but in small enough numbers to be able to identify one another sufficiently to try undercutting one another with pricing and output strategies. Oligopolists also have the opportunity to collude with one another to the detriment of the consumers of their products, but such coalitions tend to be unstable over longer periods of time.

Yet other products are characterized by very high substitutability with one another in the eyes of consumers while they still retain enough distinctiveness in consumers' eyes for some of them to be willing to pay a bit more for one variety than another. These are the goods we discussed in section 3.11 under the concept of variety and differentiated goods. The type of industry structure that produces them has characteristics both of perfect competition and monopoly (or somewhat more properly speaking, oligopoly) in the sense that producers retain some modest price power. It is, accordingly, called monopolistic competition (section 4.5). It is reasonable to think of producers of fine pottery in Archaic and Classical Greece, and even terra sigillata producers in the Roman Imperial Period, as having operated in such an industry structure.

Contemporary monopolists, as we all know from the news, generally deny their market power because most contemporary legal systems try to outlaw them. However, sometimes the claims of large firms that they face considerable potential competition are legitimate: a firm can reach a fairly large size without being able to raise sale prices to consumers much at all above their production costs. The relatively recent theory

of contestable markets (section 4.6) focuses on characteristics of fixed costs and uses the concept of the sunk cost – a cost that once expended, you can't get back by selling the thing to somebody else – to illuminate the consequences of such potential competition.

We turn in section 4.7 to the situation in which a buyer of some good or service is large enough relative to the sellers to have some ability to squeeze a lower price out of them. This situation is called monopsony and is parallel, in many ways, to monopoly for sellers. Governments frequently reach such dimensions as does the proverbial company town. Of course, the intermediate situation of several large buyers presents the parallel case of oligopsony, but most of the characteristics of oligopsonistic buyers emerge from an examination of monopsony, so we don't treat that case explicitly. It is easy to think that a monopolistic seller must be a monopsonistic buyer also, but frequently monopolists use the same array of inputs that everybody else in the economy does, leaving them little or no price power in the markets for most if not all of their inputs. In the final two sections we collect concepts from the technical sections of this chapter and apply them to the third installment of the economics of Mycenaean vases (section 4.8) and the use of competition concepts to ancient religious change and to ancient foreign trade restrictions.

4.1 Perfect Competition

In perfect competition, no individual supplier (firm or individual) believes that he can affect the price of the product he makes. Stated alternatively, each supplier believes that he can supply all he can or wants to at the current price. (The demand curve looks flat to the perfectly competitive supplier.) Only constant-returns-to-scale technologies are consistent with such expectations of costs. For such price expectations to exist, there must be a large number of both suppliers and demanders (demanders, too, otherwise some "big" demanders could cut deals with some

suppliers, and the same price would not prevail for all suppliers). The buyers must all look the same to the sellers, or they could cut special deals with (or impose special terms on) some buyers, and sellers must look the same to buyers, or sellers once again would be able to charge different prices to different buyers. Suppliers must be perfectly free to enter and exit the production of the good in question. The commodity must be homogeneous or price differences would arise to compensate for the differences in the product. Finally, both buyers and sellers must have perfect information about prices. This set of assumptions probably sounds too pristine to have ever been true, and it is – but remember, it's a model, in which the ground thickets are cleared out of the observational world so that the most important forces can be observed clearly. Letting the thickets grow back in applications generally doesn't materially affect the forces at work.

The marginal revenue for a perfectly competitive producer is simply the price of the product he sells. Since the quantity he sells does not affect the selling price, marginal revenue is constant over any quantity of the individual seller's sales: $MR = p$. Maximizing revenue involves simply maximizing sales volume minus costs: $\text{Max } R = pq - c(q)$. The first-order condition for maximizing the problem yields the condition that $p = MC$ (price equals marginal cost).¹

Perfectly competitive suppliers will earn zero profits in the long run. Any temporary profits a competitive supplier will make will be bid away by the entry of new suppliers. Perfectly competitive suppliers will produce at the intersection of their marginal and average cost curves, and the product price will just cover their average cost. If a perfectly competitive producer finds himself producing at a marginal cost above the selling price, he will reduce the quantity of fixed factors he employs so as to get his costs back down to a level compatible with non-negative net revenue (to avoid using the term "profit"). The producer in Figure 4.1 will produce and sell quantity q^* at price $p^*(= MC = AC)$ in Figure 4.1, making no profit. If there are some restrictions on entry into an otherwise perfectly competitive industry,

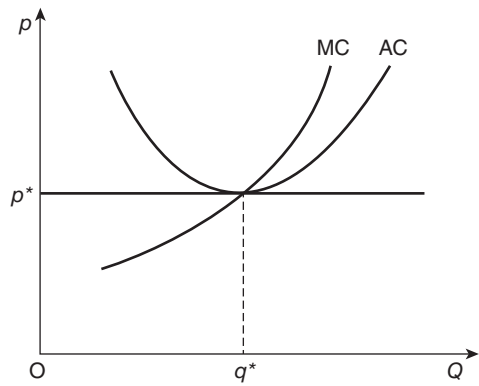


Figure 4.1 Output determination in perfect competition.

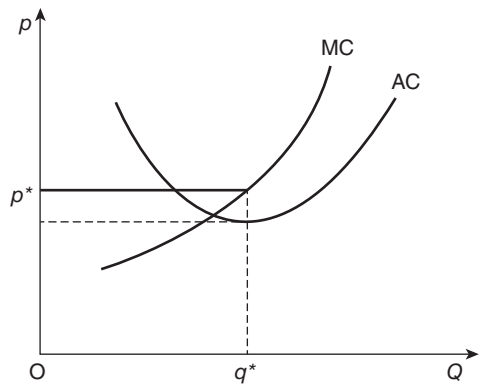


Figure 4.2 Some producers produce more than equilibrium amount at price p^* .

some or even all firms could produce on the segment of the marginal cost curve beyond its intersection with average cost, as in Figure 4.2, making what appears to be a profit equivalent to the rectangle shown in that figure. However, that profit is exactly equal to the rent accruing to the fixed factor; the producer could rent out that scarce factor to another producer and collect that much payment, so the rent is actually a factor cost to him. He does indeed "collect" this payment, but he does so as an owner of the factor instead of as the producer of the product. The profit, in an economic sense, is still zero, since he must

account the income he gets from the fixed factor as a factor cost to the production enterprise.

Competition under the conditions of perfect competition is quite impersonal. Sellers do not try to beat out their competitors by underselling them – they'll go broke quickly if they try. They know equally well that they cannot raise their prices to any of their customers because they'll lose all their sales equally quickly. The best method of competing in this environment is to focus on one's own production – maintaining productivity and product quality. What is frequently thought of as active competition – trying to undercut competitors – is actually an indication of the lack of perfect competition; to keep the industry structures straight, the latter is called rivalry or rivalrous competition.

It is clear that the information and product homogeneity requirements of perfect competition are seldom met in practice. Imperfect information – imperfect because it is costly to assemble and to use – characterizes virtually all social conditions, and a modest degree of imperfect information will not render the concept of the perfectly competitive industry useless or irrelevant. Similarly with product homogeneity, although the deliberate introduction of quality and variety differences can be studied more incisively in the context of a different industry structure somewhere between perfect competition and monopoly / oligopoly, called monopolistic competition, which we will discuss briefly below. Consider an example of a departure from perfect information. On the one hand, farming is considered the perfectly competitive industry *par excellence*: large numbers of producers, large numbers of buyers, roughly constant returns to scale, many by-and-large homogeneous products. However, imperfect information bears heavily on the supply side of this industry: farmers must commit themselves to input decisions before they have information about the weather. Their supply decisions at the beginning of the crop season are *intended* or *desired* supply decisions; were they able to forecast the weather perfectly, they generally would pick different sets of inputs. Consequently, price fluctuations are

endemic to agriculture, which has fostered the development of numerous devices for stabilizing consumption, if not necessarily current supplies or prices.

Rathbone's (1991, 369–370) assessment of the rationale for and use of the accounts found in the Heroninos Archive of the Appianus estate in third-century C.E. Egypt makes a far-reaching point: "Their [the accounts'] concern, however, seems to have been far more with the one crucial element of profitability which the estate was most able to influence, namely the cost of production . . ." A firm in a competitive industry is unable to influence the price of its product, but it can produce more or less efficiently. The implication of Rathbone's conclusion is that the Appianus estate, despite its size, produced for sale in a group of industries – viticulture, wheat, barley, olives (213–217) – which were pretty close to perfectly competitive, and used its accounting system to try to stay as close to its production possibilities frontier as it could.

4.2 Competitive Equilibrium

Although each producer / seller in a perfectly competitive industry has only an imperceptible effect on the product price, all the sellers together will affect the equilibrium price, a concept we have not introduced until now. For an equilibrium in the market for any product, the quantity demanded must equal the quantity offered, and the product price is the means of bringing these two schedules of wants and offers into equality. In equilibrium, $D(p) - S(p) = 0$, where we simplify the algebraic representations of the demand and supply curves to be functions of the one variable that they have in common, the product price. Figure 4.3 shows these curves and the price that equilibrates this market (equalizes supply and demand).

Consider fuller representations of the demand and supply functions as $D_i(p_i/p_n; p_j/p_n, y/p_n, \Theta) = S_i(p_i/p_n; p_k/p_n, p_m/p_n, \Lambda)$, in which D_i and S_i are the demand and supply functions for good i , p_j represents the price of any good or

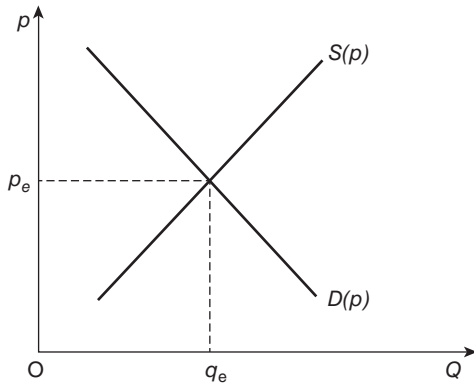


Figure 4.3 Supply and demand in competitive equilibrium.

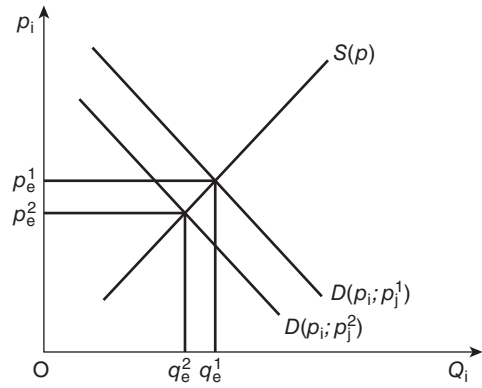


Figure 4.5 Indirect price effects on demand.

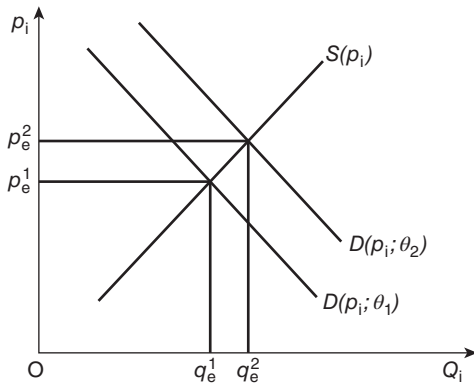


Figure 4.4 A shift in demand.

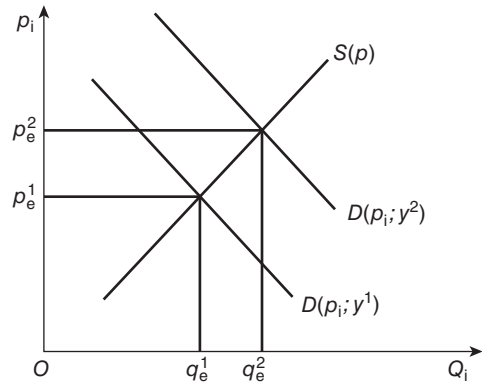


Figure 4.6 Effect of income increase on supply-demand equilibrium.

goods that may be a close substitute or complement in consumption, p_n is the numeraire good (to avoid using money as the numeraire), y is income, p_m represents goods that may be close substitutes or complements in production, and Θ and Λ are other variables that affect the demand and supply for good i independently. Figure 4.4 shows the effect of a change in the “independent demand shifter,” Θ , from a value of Θ_1 to Θ_2 . The movement of the demand curve from $D(p; \Theta_1)$ to $D(p; \Theta_2)$ is called a “shift” in the demand curve, to distinguish the change from a movement *along* the curve caused by a change in p (p_i/p_n) alone – the “own-price.” Figure 4.5

shows that a change in the equilibrium price of good 1, and the equilibrium quantities demanded and supplied, need not be induced by anything directly to do with that good. The price of good j changes from p_j^1 to p_j^2 , which has the effect of shifting the demand curve for good i downward. Is the change in p_j an increase or a decrease? It depends: if good j is a substitute for good i , we know that p_j fell; we also know that if good j is a complement to good i , p_j rose. Income changes also will shift the demand curve. In Figure 4.6, income rises from y^1 to y^2 , and the demand curve shifts outward, resulting in a higher equilibrium price and quantity transacted.

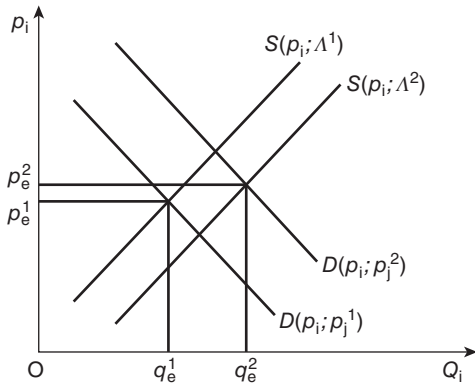


Figure 4.7 Simultaneous shifts in supply and demand curves.

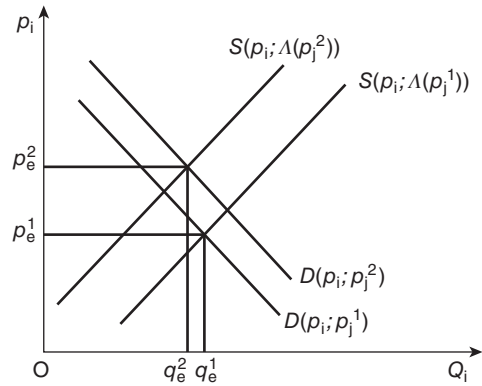


Figure 4.8 Alternative impacts of price changes on supply and demand curves.

Simultaneous exogenous changes can affect the supply and demand curves independently to give two components of change to equilibrium price and quantity. Figure 4.7 shows a change in the price of the substitute/complement to good i shifting the demand curve outward and a change in the supply shifter, Λ , shifting the supply curve outward at the same time. The increase in supply is insufficient to keep the price constant in the face of the increase in demand.

It is possible that the same exogenous influence can shift both supply and demand at the same time. We have not specified such a common variable (note that p_i/p_n is not a “shifter” variable, but is the independent variable in whose “space” (the vertical axis) the supply curve is defined). We could have made the supply-shift variable Λ a function of the price of the substitute/complement good in consumption, p_j/p_n : $\Lambda = \Lambda(\Lambda^*, p_j/p_n)$. Then in Figure 4.8, the change in p_j shifts out the demand curve and shifts back the supply curve. Why? Suppose that good j is a substitute for good i . Then its price must have fallen. This same rise in p_j caused a change in the value of the supply-shifter function, Λ , that increased the cost of supplying any quantity of good i , since the supply curve has shifted back toward the p_i axis.

4.3 Monopoly

A monopolistic producer / seller affects the price of her product by her sales volume. She does not

take the sales price as given. Consequently the marginal revenue from each subsequent unit sold is lower than that for the previous unit. Thus, the monopolist starts with the same revenue maximization problem that the perfectly competitive supplier did above [$\text{Max } R = p(q)q - c(q)$], but in her case, her sales price, as well as her costs, is a function of her output, q . This revenue maximization problem does not yield the recommendation to price one’s output at its marginal cost, but at something possibly considerably higher than marginal cost. The first-order condition (and all the little variations of q to see which value of q gives the largest net revenue) for profit maximization under monopoly tells the monopolist to set $\text{MR} = p(1 - 1/\epsilon_d)$, in which $\epsilon_d > 0$ is the negative of the elasticity of demand for the product. Marginal revenue will also equal marginal cost, so the relationship between price and marginal cost for the monopolist is $p = (1/\epsilon_d - 1)\text{MC}$, which gives a profit-maximizing selling price that could be substantially in excess of marginal cost. When the value of ϵ_d goes to infinity (what the “market” demand elasticity looks like to a perfectly competitive supplier), marginal revenue coincides with the demand curve and price equals marginal cost. The smaller is the demand elasticity, the greater is the “monopoly power” of the seller, in that she can raise price further above marginal cost.

Figure 4.9 shows the price and sales-quantity determination of the monopolist. The demand curve facing her is labeled D . Her marginal revenue curve is MR . Marginal and average

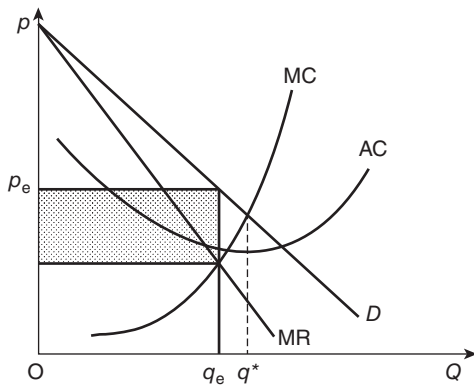


Figure 4.9 Equilibrium in a monopolistic market.

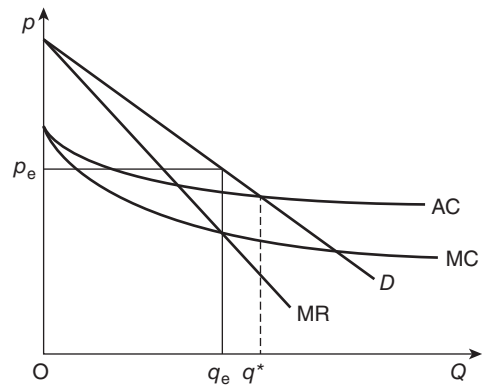


Figure 4.10 Profitable additional sales for a monopolist.

cost are MC and AC. To maximize profit, she picks the sales quantity (q_e) that will equalize marginal revenue and marginal cost, but that $MR = MC$ intersection is considerably below the maximum price that consumers will pay for that quantity of output: read up to the demand curve directly above q_e and the $MR = MC$ intersection to find the equilibrium, profit-maximizing price, p_e . The area of the shaded rectangle is the monopolist's "profit."

Characteristics of the production technology are the most common source of monopoly. High fixed costs or increasing returns to scale can result in average and marginal cost curves that are still falling in the region where they intersect the industry demand curve (recall the discussion around Figure 2.12). This condition leads in turn to lower cost production by a smaller number of larger suppliers rather than by the indefinitely large number of small suppliers characteristic of constant returns to scale and perfect competition. Monopoly technically consists of a single seller in an industry (hence the "mono-" prefix), and cost competition may indeed drive out all but a single supplier unless stopped by government intervention.

Monopoly "profits" are actually a species of rent rather than a true profit as a reward for successful risk taking. If a monopolistic firm were to be sold, the original owner would be able to command inclusion of the value of the rents into the sale price, and the second owner – and all subsequent owners – would earn only the competitive rate of return on their capital.

The inefficiency of monopoly pricing can be seen in Figures 4.9 and 4.10. In both figures, although the monopolist maximizes her profits at price p_e and quantity q_e , consumers would be willing to pay p_e , or just a tiny bit below it, for the next unit or so of the product, and the monopolist could still make a profit from selling those units, up to the point where the marginal cost curve cuts the demand curve, which would give the $p = MC$ rule for pricing. Unless a separate pricing plan can be arranged for the units of output between q_e and q^* , they will not be supplied to people who would be willing to buy them at prices at which they could be profitably sold. In Figure 4.10, in which AC and MC are still falling, with MC below AC, the profitable area of continued sales by the monopolist would be restricted to the q^* where demand equals average cost.

As another source of monopoly power, government may bestow monopoly rights on a firm. On the other hand, in such a bequest, the government may place requirements on the firm that sap the profitability of the original rights. For instance, a firm might be given the sole right to produce, say, bronze implements in a country, but be required to supply the army with bronze weaponry at prices below its profit-maximizing level.

Notice that we have not drawn a supply curve in the analysis of monopoly behavior. A monopolist is generally considered not to have a supply curve because there is no particular quantity of output it would want to offer at particular prices. Its output decisions determine price, in conjunction

with the industry demand curve. Monopolies need not be, and generally aren't, permanent, as technologies, or even tastes, may change and erode or even obliterate their advantages.

4.4 Oligopoly

The other major source of monopoly profit may be collusion, but this is a weaker and more tenuous source of such rents. The underlying technological conditions that permit successful collusion exclude perfectly competitive behavior. They will result in several suppliers being able to make profits in the same market, but not enough suppliers to totally eliminate profits. Colluders may make agreements on output shares to keep prices up, but each partner to the collusion has incentives to cheat. As the number of colluders rises, the difficulty of identifying cheaters increases, even though all partners to the collusion may be able to tell that someone is cheating. Different partners may cheat at different times and uphold the agreements at other times, adding to the difficulty of detecting cheating.

The analysis of oligopoly is facilitated by looking at the case of two rival firms, the case of duopoly. A given number of customers (translating into a fixed quantity of demand) can be served by the two firms. Each firm understands that the price they both receive depends on the quantity of their products they both produce. Consequently, each must make some planning assumptions about the interactions it will have with the other firm. To study this planning generally, we can think of each duopolist maximizing its net revenue, but finding some way of taking account of the influence of the other firm's supply behavior on the product price: $\text{Max } p(Q)q_i - C(q_i)$, where Q is the output of the entire industry (two firms since we are thinking in terms of duopoly). Each firm's first-order condition for maximizing profit is that $p(Q) + (\Delta p(Q)/\Delta q_i)q_i - \Delta C(q_i)/\Delta q_i = 0$. When it tries out (in its planning room, not in practice, of course) different values of output, q_i to see what output will maximize its profit, it first isolates the price as a first approximation, then subtracts the effect that its own output would have on the

market price by adding to the total supply, Q , then looks to how its production decisions affect its cost. At the value of q_i for which the sum of these three effects equals zero (makes no further change in net revenue), it has found the output that will maximize its profit.

But this still has not identified for us how each firm thinks about the interaction with the other firm. We can study that decision by rearranging slightly the price-change term in the first-order condition, rewriting $\Delta p(Q)/\Delta q_i$ as $(\Delta p(Q)/\Delta Q)(\Delta Q/\Delta q_i)$. The expression in the first set of parentheses is the effect on price of the change in total industry output, and that in the second set of parentheses is the extent to which the firm's own output changes industry output. The direction and even the size of the first effect – the effect of industry output changes on the industry price – is of lesser concern, both to us as students and to the firms as actors. The second effect, however, embodies the firm's expectations of reactions by the rivals, as well as the impact of that reaction on the industry price. If it believes the value of that term to be zero, the firm effectively considers itself a perfect competitor: its output has no perceptible influence on industry supply and consequently no perceptible effect on industry price.

If the firm believes the value of that term to be 1 (or alternatively, if we as students of oligopoly consider it to have the value of 1 so we can study the implications of such behavior), the firm believes that its rival will not change its output in response to its own production decision; it is planning on accepting the full reduction in industry price that its own supply decision would bring. This anticipation is known as Cournot–Nash behavior: each firm supposes its rival's output to remain fixed when determining its own supply decision. With homogeneous products, each also understands that if it raises its price a small bit above its competitor's price, it will lose all its sales – and that if it lowers its own price a small bit below the competitor's that it can capture all the sales. Since they both understand this, they will both be forced down to pricing at marginal cost. If we assign the value of Q/q_i for that term, we have collusive behavior, in

which firms try to find a set of sales shares that is agreeable to all rivals (going beyond the duopoly situation for the moment).

A more sophisticated anticipation that a firm could have of its rival's behavior would be embodied in letting the $\Delta Q/\Delta q_i$ term have the form $\Delta Q(q_i)/\Delta q_i$, which has the interpretation of the industry's particular supply reaction to this particular firm's supply decision. That is, the firm forms an assessment of how either the other rival firm or the rest of the industry (stepping away from duopoly again) will alter its supply in reaction to its own move. It is this firm's assessment of how its rival(s) will respond to its action. Different anticipations (expectations) of rivals' actions can lead to different actions in the first place to maximize profit. This is called Stackelberg behavior.

More generally, noncooperative game theory can be used to model the strategic interactions of oligopolistic firms. These tools let the analyst specify what information each firm has about the other and when it gets it. In this type of analytical framework, each firm knows what actions it can take under various circumstances (at each "move" or "play" of the "game"), and it will know what the other firm's options are, but it will not necessarily know what that firm's information is, so it will not know for sure which "path" its own choices may be taking it down, because the path down which the strategic interactions progress depends on both firms' actions, which in turn depend on the firms' knowledge sets and their anticipations (best guesses) of how the other firm will behave. While a firm may not be able to predict exactly how its rival(s) will behave, to the extent that it understands the structure of the game (the actions each firm can take from the beginning of the interaction through its end point), and the payoffs (rewards, profits) of the rival(s) in alternative end-states, it can make estimates of the likelihood of what strategies its opponents will pursue. Of course, the rivals are making the same sorts of assessments about the firm in question (the game theoretic models are symmetric across "players"). When one of these interactions (games) occurs only once, firms don't have much opportunity to develop an understanding of how strategies may be varied to affect payoffs to the mutual advantage of each player. In repeated games, this opportunity becomes available, and

even the famous Prisoner's Dilemma game can be played so both strategists win rather than lose. The use of game theory is becoming widespread and important in the study of industry structure and behavior as a device to study interactions under limited and asymmetric information, but its mathematical demands are generally considerable (see, for example, Fudenberg and Tirole 1991; Myerson 1991).

4.5 Monopolistic Competition

The monopolistic competition model of an industry structure is sort of a "half-way house" between the perfectly competitive industry in which the implicit competitive pressures among a large number of sellers keep the prices of identical products exactly equalized across sellers and the monopoly model in which a single seller's product is sufficiently unique that he pays no attention to the prices of other producers. In reality, many, if not most, products have some degree of differentiation, such that sellers have some small degree of ability to pick their product's price without being run out of the market altogether and immediately. The typical firm has to pay some attention to the prices of competing products that consumers might substitute for their own, should their price get too high. The monopolistic competition model captures these features – large number of sellers and imperfect substitution among the differentiated products of different sellers – and offers as its analytical results information on the equilibrium number of firms, price and output.

Figure 4.11 shows the downward-sloping demand curve the monopolistically competitive firm perceives for its product, its marginal revenue curve, which reflects the fact that it has some influence over the price of its differentiated product, and its average and marginal cost curves. Its maximization problem is $\text{Max } p(q; q_i)q - c(q)$, which differs from the monopolist's by virtue of the fact that the seller has to pay attention not only to its own output, q , but also to the entire array of outputs q_i offered by i different firms. Consequently its first-order condition for profit maximization contains more than the requirement for it to equalize price and marginal cost, as a perfect competitor would, and not exactly to set

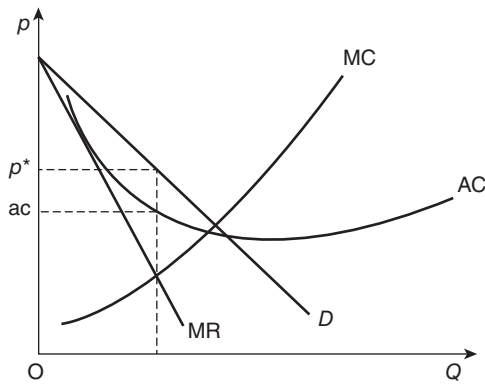


Figure 4.11 Monopolistic competition.

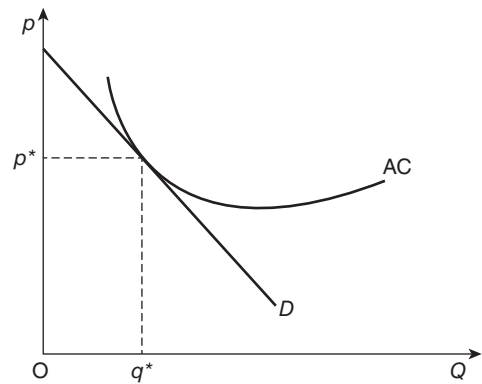


Figure 4.12 Equilibrium in monopolistic competition.

price equal to a mark-up above its marginal cost related to the elasticity of demand for the product, as the monopolist does. Instead, it sets its price equal to its marginal cost plus a factor that takes account of its own pricing policy on the ability of competing, differentiated products to erode its own sales. As other firms enter to capture some of the temporary profits, the monopolistic competitor finds its demand curve shifted to the left. This competition from other firms' products erodes the area of profits shown in Figure 4.11 until they completely disappear, during which process the monopolistic competitor's demand curve becomes just tangent to its average cost curve, resulting in the equilibrium pricing policy that price is set equal to average cost rather than marginal cost, as shown in Figure 4.12. This was long thought to involve some degree of inefficiency because this tangency must occur on a downward-sloping portion of the average cost curve rather than at the minimum, but the recent attention to the benefits of product variety has altered that conclusion. These firms producing relatively small quantities of goods whose exact replicas are not produced by any other firms simply do not face sufficiently great demand to exhaust the potential economies of scale in their production processes. This accounts for their production at points on the downward-sloping part of the average cost curve but does not imply that the quantity demanded could be produced at lower cost.

The theory of monopolistic competition has existed in the economics repertoire since 1933, but only recently has it begun to bear analytical fruit. Most economists would have agreed that its description of firms having some degree of influence over their own prices, but being subject to possibly intense competitive pressures, was descriptively more accurate than either the perfect competition or monopoly models but it offered no analytical predictions that were not offered more clearly by those other models. Since the mid-1980s, more useful results have begun to be derived from the monopolistic competition model, particularly using some of the newer approaches to product differentiation discussed in Chapter 3. I include this much discussion of the monopolistic competition model more for the sake of completeness than as an implicit recommendation regarding its utility in illuminating the behavior of firms in the ancient world. Some students may find it worth pursuing in the ancient context.

4.6 Contestable Markets

The theory of contestable markets shifts attention from actual competition – the presence of several firms in a market – to potential competition – the readiness of firms to enter a market should conditions appear attractive. The traditional theory of

monopoly has focused on fixed costs and increasing returns to scale (which may be attributable to the existence of fixed costs) as the barriers to entry that permit a monopolist to remain in a market by itself. The concept of fixed costs is a subtle one that depends on the length of time the producer incurring them *intends* to be using the assets called fixed and what proportion of their remaining value (after depreciation) can be obtained for them in an earlier-than-planned sale.

The notion of the sunk cost accompanies that of the fixed cost and has to do with the retrievability of an investment required for some line of production. A sunk cost is a fixed investment cost that cannot be recouped through sale of the asset; it can only be recovered by continued production, if otherwise warranted. There is an old adage in economics that sunk costs are “bygones.” They do not affect current production decisions because you’ve already incurred the cost and nothing you do can either get it back for you or really make it worse – unless you throw more money at it. The kind of fixed costs that are sunk costs would be highly process- or product-specific machines, possibly ones made especially to order for your very own, particular production conditions; investments in training labor for work that is not transferrable to other operations of the employer; and so forth.

If fixed costs are sunk costs – the question really is what proportion of fixed costs are sunk – an incumbent in a market (the firm already producing in the market) – has a production cost advantage over firms looking at entering the market who haven’t paid their sunk costs yet. From the prospective entrant’s perspective the fixed costs aren’t sunk yet – they’re truly variable. From the perspective of the incumbent firm, they don’t even exist. Consequently, even with identical technology, incumbent firms have lower unit costs than do potential entrants when there are sunk costs.

However, when sunk costs are minimal, even if “reversible” fixed costs are substantial, we have a condition that has been called a “contestable” market, one in which the competition from potential entrants affects the market behavior of incumbent firms. In markets with this characteristic, firms will price and produce according to the intersection of demand and average cost curves, giving them zero profits even though

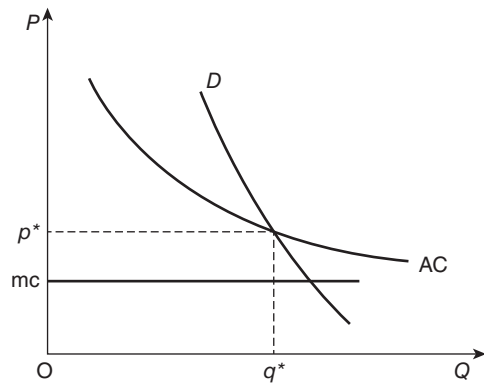


Figure 4.13 A contestable market.

their marginal cost curves are below the average cost curve. The average-cost pricing is efficient as long as the firm is offered no subsidy to cover the difference between the average and marginal cost. Figure 4.13 shows this. D is the demand curve, AC is the average cost curve, and mc represents a constant unit marginal cost. Price p^* is the lowest price the firm can charge and keep profit non-negative. A monopolist would produce further to the right, at the intersection of mc and D , charging a slightly lower price but earning considerable monopoly rent.

4.7 Buyer's Power: Monopsony

We have mentioned monopsony in passing already. It is an alternative to the competitive model of the purchasing behavior of an industry. Under perfect competition, firms take the prices of their inputs as given to them, although if they all were to attempt to secure more of any particular factor, they could drive up its price. The profit maximization problem of the perfectly competitive industry was $\text{Max } p_q \cdot f(x_i) - p_{x_i} x_i$, which led to the first-order condition that said that $VMP_{x_i} = p_{x_i}$: the firm needs to employ each input x_i in an amount that will equalize the value of its marginal product to its price. The monopsonistic industry, being the only demander of some particular input it uses, affects the price of those factors by its employment decisions. We represent its consequent profit-maximization problem as $\text{Max } p_q \cdot f(x_i) - p_{x_i}(x_i)x_i$, which indicates that the prices of at least some of its inputs

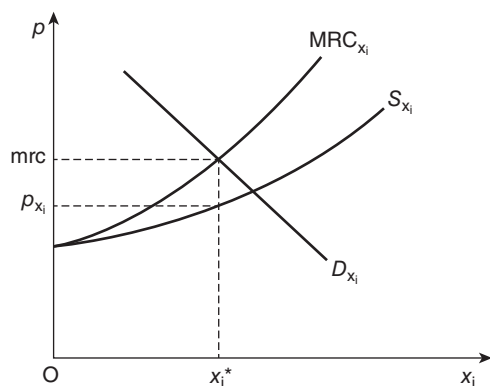


Figure 4.14 Monopsony.

are functions of the quantity it demands. The first-order condition for this setting is parallel to the first-order condition for the monopolist's product pricing policy: the monopsonist employs a quantity of the factor over which it has price power so that its value of marginal product equals the factor price times the sum $(1 + 1/\epsilon_s)$, where ϵ_s is the supply elasticity of the factor in question.

Figure 4.14 shows a monopsonist's demand for an input, x_i , a supply curve for that input, and a curve marginal to the supply curve, as the marginal revenue curve was "marginal" to the monopolist's demand curve. This curve is called the marginal resource cost, and it indicates the change in the cost to the firm of the factor in question, just as the marginal revenue curve indicated the change in total revenue caused by the fact that the monopolist's output affected its product price.

4.8 The Economics of Mycenaean Vases, III: Industry Structure

Two salient facts of Mycenaean pottery are that some greater degree of standardization (a *koiné*) emerged by Late Helladic IIIA2, after considerable diversity in LH I and II (Mountjoy 1993, 5–15); and the volume of Mycenaean vases exported to Cyprus and the Near East (including Egypt) reached a peak in LH IIIA and B (Mountjoy 1993, 163–172). We'll return to these twin facts.

Pottery production involves no considerable fixed costs or other source of increasing returns to scale that would place its market structure

into either monopoly or oligopoly. Perfect competition or monopolistic competition are the most appropriate models to characterize pottery markets. Perfect competition likely would characterize the supply of plain and utilitarian pottery, while variation in shapes, decoration and quality would support the use of the monopolistic competition model in analyzing the supply of fine pottery. In either case, producers make no economic profits (they just cover the required return on capital, including equipment and skills), and selling price equals either marginal or average cost. This equality justifies the practice of examining the sources of production costs to approximate what sale price would have been.²

We can safely ignore, by and large, transportation costs of pottery sold locally (unless, of course, that is our focus of interest – but archaeologists have expressed little interest in such local transportation costs of pottery). Transportation costs on pottery sold to foreign, overseas markets generally shouldn't be ignored, because of their likely magnitude relative to "factory" cost. We'll use the terms f.o.b. (freight on board) and c.i.f. (cost, insurance, freight) prices to refer to prices at the door of the pottery and the delivered price, including all transportation costs, at an overseas market. The cost components in pottery production we discussed in Chapter 2 refer to f.o.b. costs. It is quite likely that a shipment entirely of fine pottery would have been able to bear shipping costs³ but it's possible that the marginal cost to a carrier⁴ of small and irregular shipments of pottery could have been zero, making the "f." part of the c.i.f. price small or occasionally vanishing. Nonetheless, the shipper still would have to pack these vessels carefully, possibly in seaweeds such as eel grass, and then inside shipping amphoras for the smaller vases, to avoid breakage. Additionally, experience would have taught either shippers or carriers to include a charge at the destination for breakage en route: the equivalent of a self-insurance fee.⁵

With regularized shipments, the "f." component might come under a long-term contract with one or more carriers, with the tariffs (the shipping costs or rates) becoming more-or-less public knowledge. Thus, the LH I and II vases in the Near East might have irregularly incurred freight charges of any magnitude, although they still would have had to be packed well. By the

time of the LH III ceramic export boom (if that isn't an exaggeration), we should expect freight charges to have been incurred regularly.

Supposing that the cost of producing fine vases in the Argolid was roughly what it was in the hinterland of Ugarit, we can reasonably consider that the c.i.f. prices of Mycenaean vases in Ugarit were higher than the c.i.f. prices of local vases of similar dimensions and shapes. It is also possible that the transportation costs on the Mycenaean vases might have offered the equivalent of a tariff protection to the local, Ugaritian products, allowing their sale prices to rise to a level comparable to their imported competitors. In this case, the Mycenaean articles wouldn't have cost much more than the local varieties, although the local potters – or middlemen, or somebody in the distribution system – would have collected the difference between the local production cost and the sale price.

We might think also of the steps by which some regular price for Mycenaean vases became established in Ugarit. When the first shipments of Mycenaean vases arrived there, their novelty well might have commanded a considerable premium. Whoever sold them for that price – the producer's representative, the ship's captain, or the merchant – would have thought to himself, "Hmmm.... I could get rich at this rate. I'll go back to Mycenae and buy up everything old Eleutherios has and bring it back here!" He does that. Asks Eleutherios for his entire stock – or actually just the good stuff. Eleutherios actually had half of his inventory already earmarked for several local distributors⁶ and it took an extra consideration to persuade him to disappoint them. But the enterprising merchant took them even at the higher price (he was working up a supply curve) and took them to Ugarit, where he discovered that he was working down the demand curve there when the unit prices at which he was able to move the larger load were lower than before. Maybe he still made a profit, maybe not. At any rate, the information generated in the second round of sales (in actuality, it could have taken several iterations and several different agents involved in the export-import business) would have begun to establish an expected price range. F.o.b. prices in Mycenae could have been driven up by the increased demand from Ugarit and the rest of the Canaanite coast, but the

elasticities of substitution of the customers in Ugarit – between their local varieties and the Mycenaean varieties – would have placed an upper limit on both what they would have been willing to pay and how many the Mycenaean potters would have been able to produce.⁷

When some internationally traded goods are the sort of differentiated goods supplied by monopolistically competitive firms, the opening of trade will reduce the number of such firms (and correspondingly their varieties) and increase their typical size. This theoretical result brings us back to the coincidence of the timing of the increased standardization of Mycenaean decorated pottery and the peak in its export volume. Could it be possible that the emergence of the Mycenaean *koiné* of fine, decorated pottery from the exhilarating variety of the earlier Late Helladic periods was a result of the opening of trade and the expansion in size of the typical pottery production unit?

4.9 Ancient Monopoly and Oligopoly: Religion and Foreign Trade

In a natural monopoly, increasing returns to scale are such that the optimal size of producer becomes so large that only one producer can serve the market. Money may be about the only example of natural monopoly from the ancient world. Legally created monopoly, in which only one sanctioned producer is allowed, probably was about as prevalent in the ancient Mediterranean societies as it has been in recent – and contemporary – times. In either type of origin, sellers are able to charge prices higher than their production costs; while they may have to share the profits with the agent offering the legal sanction, they probably still retain some profit. Scholars have believed, on varying types of evidence, that foreign trade was so conducted at various times and places, in the ancient Near East particularly, but by Rome as well. The concept, or at least the metaphor, of monopoly has also been applied to religious coercion in the ancient world, with substantial intuitive appeal. We explore the monopolization of religion first.

Scholars have appealed to the concept of monopoly in discussions of ancient religions,

particularly in “elite” manipulations of religion to acquire, enhance, and maintain political authority and power. The notion is intuitively appealing, since something about the “control” of religion sounds calculated to confer power and authority on whoever does the controlling. Explanations along these lines have been offered for the transition in Crete from the Old Palace Period to the New, coincident with a noticeable reduction in dispersed, rural peak sanctuaries, probably around 1750/20 B.C.E. (Cherry 1978; Peatfield 1987). The argument would have the new rulers consolidating their authority by monopolizing religion, which is reflected in the curtailment of peak sanctuaries and the enhancement of religious sites close to the palatial city of Knossos, particularly the peak sanctuary on Mt. Juktas, just to the south. Peatfield (93) used the analogy to the Church of Rome’s efforts in the first century C.E. to monopolize the supply of Christianity in Western Europe as a parallel to the Minoan episode. Let’s explore this application of the monopoly model.

In “official” attitudes toward religion in the eastern Mediterranean lands – particularly Egypt and the great kingdoms of Mesopotamia, but possibly in Bronze Age Crete and Greece as well – the ruler typically assumed a key intermediary role vis-à-vis the deities, either as the chief among priests or a quasi- (or wholly) deified link between the people and the deities. While this role involved restrictions on who could occupy certain offices and undoubtedly conferred secular power as well as religious, it is not really a case of monopoly, as we have treated that model: restricting the supply of a good or service so as to drive up its price. Are there other routes by which these rulers genuinely could have monopolized religion?

Consider some possibilities of what the monopolization of religion might have looked like in Minoan Crete. We probably should look to a combination of legal sanctions on the practice of other religions – an artificial supply restriction – and palatial supply of the legally approved alternative: banning of activities at all but a few peak sanctuaries and channeling of all rituals through those few sanctuaries. However, the analogy

to “raising the price of religious services” by restricting their supply is not obvious. In the first place, a former religion, or at least some practices, may have just been outlawed, eliminating the “payments” worshipers may make for those services. Second, if new practices are substituted for the old, outlawed ones, their supply increases rather than decreases. The amount of “payment” worshipers would be willing to make for these new services (rituals) would depend on the substitutability of the new services for the outlawed ones. However, assessment of the pattern of types of votive offerings at peak sanctuaries during the Old and New Palace Periods shows a good bit of continuity (Jones 1999, esp. 76, Table 12). Accordingly, to the extent that the votives reflect the religion, the substitution of one religion for another doesn’t receive a lot of support from the archaeological evidence. The contraction of the number of locations at which a relatively unchanged religion, or set of rituals, were offered certainly would restrict the aggregate supply of those religious services, but the pattern of restriction facing the consumers would not be well calculated to implement a monopolistic supply restriction: some consumers (many, in fact, considering the dispersal of the population across the island) would have become effectively unserved altogether while those living near the continuing sites might actually have had their supplies increased. Consequently one group (or many groups) of consumers of the religious services would have had their “payments” effectively terminated through termination of supply while another group would have faced no noticeable supply restriction.⁸ The mechanism for squeezing greater payments of one sort or another out of consumers by working back up their demand curves seems to be missing.

How would the people have “made the payments” and who would have received such payments? Votive offerings could have been a means of payment, and shrines, a priesthood, the gods (or all the above) may have been the recipients. How could the ruler have benefitted? Taxation could have been simpler than passing through votive donations and converting them into something the ruler could actually use.

Alternatively, propitiation of the ruler's deities by getting people to worship more than they would have otherwise is a likely source of benefit to the ruler but it is not clear that restriction of people's access to religious services is the mechanism used to increase the "price" they pay (the "revenue" going to the ruler, in the form of votive donations, increased favor with the gods, or both). So, while methods of making "payments" for religious services can be specified with little difficulty, it still is not clear that the particular method of proposed supply restriction would permit those mechanisms to operate to the benefit of the agents supposedly imposing the monopolistic supply restriction.

Religious systems with separate sanctuaries to multiple deities, such as characterized Archaic and Classical Greece, would have been ripe for oligopolistic behavior on the part of the different deities' priests. While the different deities were complementary to one another in many ways, they certainly offered religious services that were at least partial substitutes for one another. The different sanctuaries would have competed on the basis of differentiated products (oracles versus healing, for instance; oracles with different specialties; the famous games) as well as prices. "Prices" required for various services would have taken the form of votives and more direct payments for explicit services such as divining. It may be difficult to discern the sanctuaries' "pricing policies" from the distance of two-and-a-half millennia, but some of them may have taken such forms as the materials of which acceptable votives were made – typically bronze at one site, typically terracotta at another, wooden items being acceptable at another, and so on. The temples to various deities in Egypt well may have had comparable competition strategies vis-à-vis one another as well.

Monopolization of foreign trade offers several possibilities. First, the state could either claim the right to conduct all foreign sales and purchases itself or endow specific persons with that right. These institutions emphasize restrictions on the supply of imports to one's own countrymen, particularly if the government auctions off the right to import, a cost that must be passed on to consumers in the form of higher prices, which dampen the demand for imports. Alternatively, the government could control the quantity of exports it sells to foreigners; if the country was a relatively large supplier of the goods it exported, that restriction would increase the price of its exports. The two methods of foreign trade restriction have the same effect (reducing imports effectively reduces the ability to export, and vice versa).

Egypt's relatively cheap access to gold – relative to that facing its northern neighbors in Mesopotamia (Monroe 2005, 177, 181) – could have given it the kind of price power over gold that would have made monopolistic restriction of exports profitable. Of course, Egypt was not the sole source of gold during the Bronze Age, so an oligopoly model might be more appropriate factually. Whether Egypt and its competitors would have strategized against one another in terms of quantities supplied and price cuts is an open question. Cypriot supply of copper is another potential case of monopolistic opportunity, or more likely, oligopoly again, as the copper-producing regions of Cyprus may have been in different kingdoms, not to mention other copper-supplying regions outside the island. Each of the suppliers likely would have had some international price power in copper but somewhat less in geographical regions where transportation costs permitted several suppliers to compete effectively.

References

-
- Boardman, John. 1988a. "Trade in Greek Decorated Pottery." *Oxford Journal of Archaeology* 7: 27–33.
- Boardman, John. 1988b. "The Trade Figures." *Oxford Journal of Archaeology* 7: 371–373.
- Cherry, John F. 1978. "Generalization and the Archaeology of the State." In *Social Organization and Settlement: Contributions from Anthropology, Archaeology and Geography*, Part ii. BAR

- International Series (Supplementary) 47(ii), edited by D. Green, C. Haselgrove, and M. Spriggs. Oxford: British Archaeological Reports, pp. 411–437.
- Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory*. Cambridge MA: MIT Press.
- Fülle, Gunnar. 1997. "The Internal Organization of the Arretine Terra Sigillata Industry: Problems of Evidence and Interpretation." *Journal of Roman Studies* 87: 111–155.
- Gill, David W.J. 1988. "Trade in Greek Decorated Pottery": Some Corrections." *Oxford Journal of Archaeology* 7: 369–370.
- Jones, Donald W. 1999. *Peak Sanctuaries and Sacred Caves in Minoan Crete: A Comparison of Artifacts*. Studies in Mediterranean Archaeology and Literature, Pocketbook No. 156. Jonsered: Paul Åströms Förlag.
- Monroe, Christopher M. 2005. "Money and Trade." In *A Companion to the Ancient Near East*, edited by Daniel C. Snell. Malden MA: Wiley-Blackwell, pp. 171–184.
- Mountjoy, P.A. 1993. *Mycenaean Pottery; An Introduction*. Oxford: Oxford University Committee for Archaeology.
- Myerson, Roger B. 1991. *Game Theory; Analysis of Conflict*. Cambridge MA: Harvard University Press.
- Peatfield, Alan A.D. 1987. "Palace and Peak: The Political and Religious Relationship between Palace and Peak Sanctuaries." In *The Function of the Minoan Palaces*, Skrifter Utgivna av Svenska Institutet i Athen 4°, XXXV, edited by R. Hägg and N. Marinatos, 89–93. Stockholm: Paul Åströms Förlag.
- Rathbone, Dominic. 1991. *Economic Rationalism and Rural Society in Third-Century A.D. Egypt; The Heroninos Archive and the Appianus Estate*. Cambridge: Cambridge University Press.

Suggested Readings

- Becker, Gary S. 1971. *Economic Theory*. New York: Knopf. Chapter 6.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapters 9–12.

Notes

- 1 The producer/seller increases q "in small increments" until the additional revenue (marginal revenue) goes to zero. Increasing q in such a fashion means increasing the product of p and q (pq = revenue and $p \times \Delta q$ is marginal revenue) at the same time it increases total cost through the $c(q)$ function: $\Delta C/\Delta q$, which is marginal cost. Essentially, we have $p\Delta q - (\Delta C/\Delta q)\Delta q = 0$; rearrange this to read $p\Delta q = (\Delta C/\Delta q)\Delta q$, divide both sides by Δq , and we get the first-order condition for revenue-maximizing output, $p = MC$.
- 2 Despite archaeological evidence of some very large structures involved in terra sigillata production, Fülle (1997) casts doubt on the existence of significant economies of scale despite some modest scope (at best) for division of labor and associated specialization in larger firms (139–140). Some very large clay-processing basins (levigation tanks for clay) have been found, one of 10 000 l capacity, another of 40 000 (113, 134), but their existence does not necessarily indicate large-scale pottery production: the tanks could have been shared or could have specialized in supplying workshops elsewhere with high-quality clay (135). Kiln size also varied considerably, with one near Arezzo covering 8.16 m² while another in Arezzo covered only 1.13 m² (136). There was a large number of potteries just in Arezzo: "we know of about 110 groups with known personnel and producers without known personnel who were active in Arezzo" (133–134), which, although these were surely spread over a number of years, reinforces Fülle's contention that the evidence points to mostly small units manufacturing these pieces (134–139). What are occasionally interpreted as branch workshops by handle stamps may have been simply temporary locations of migrant or itinerant potters (141–142). Altogether, the sheer number of operations as well as the coexistence of facilities of considerably different size, strongly suggests a competitive industry in which scale economies were exhausted quickly.
- 3 See the exchange between Boardman (1988a, 1988b) and Gill (1988) on shipping costs of Attic pottery relative to its market prices in Italy in the Classical period.
- 4 That is, to the ship and its master; the "shipper" is the producer of the pottery who wants it transported someplace.

- 5 Who would have required such a fee depends on who took ownership of the vases at what point. If the producer owned them all the way to, say, Ugarit, he would have learned to add on some percentage to the asking price to cover the vases broken in shipment. If he sent them in care of an employee responsible for selling them and returning with the proceeds, that person would have had to raise the price of surviving vessels in Ugarit. If the producer sold them to either a merchant or to a ship's captain who then took responsibility for them, that individual would have to raise the price of surviving vases in Ugarit to cover his own purchase price.
- 6 Not a term identified in Linear B yet; distributors undoubtedly were more informally organized than product distributors today, but distributors probably existed during the Mycenaean period.
- 7 Because they were working up their own supply curve, which pairs up price with quantity.
- 8 The possible substitution of religious services in caves for roughly similar services on mountain peaks finds some support in the archaeological evidence but whether the agents imposing the supply restriction could have benefitted from that substitution simply isn't clear from the evidence. If the services on the peaks were actually suppressed, the substitute services in the caves could have been largely clandestine, avoiding the intended, monopolistic "taxation."

General Equilibrium

5.1 General Equilibrium as a Fact and as a Model

So far our analysis has used a partial-equilibrium framework. That means that we implicitly considered the indirect effects through other markets connected to the markets we were studying to be small enough to be ignored. An alternative interpretation of partial equilibrium analysis is that it contains the first-round effects, before ramifications through all the rest of the markets in the economy. These other ramifications can add so much complexity that it can be difficult to get any clear understanding from the results, so we start with partial-equilibrium analysis to get a first-approximation understanding that we can complicate later as we decide is appropriate or necessary.

Most partial equilibrium analysis deals with a single market or a single firm or a single household – or small numbers of them. What is omitted, but not without good reason, is the repercussions of the price changes. Changes in prices, if extensive enough, translate into changes in incomes, both wage and capital income, which are held constant in partial equilibrium analysis (the income effect in the basic consumer demand model is not a general equilibrium effect, but

rather an implicit consequence of a price change without any explicit change in income – which is why it's called an “income effect” rather than a change in income). If income changes are extensive enough, it is possible that a general equilibrium decrease in the price of a good could lead to a decrease in the demand for it. This particular example is not an especially likely effect but it is an example of the kind of surprise, relative to partial-equilibrium results, that can emerge from a general equilibrium analysis. However, overall, if general equilibrium models didn't give different results from partial equilibrium models, the general equilibrium models wouldn't have had the life span they've had.

General equilibrium theory in economics has some very abstract branches that work on important, but rather remote topics such as stability of equilibrium in replicable economies. These issues are important because they may (will) one day give deeper insights into the long-term behavior of entire economies – for example, will an economy with certain characteristics slowly “extinguish itself” – have its production and consumption (and population!) go to zero? (Remember that our sun itself is gradually consuming itself.) We won't dwell on these topics, not because they are unimportant but because

their applicability to issues important to the study of the economic lives of the ancient civilizations in the Mediterranean and Aegean is rather remote. Even the issues of the “collapse of complex societies” can be studied fruitfully with less abstract economic tools, although general equilibrium analysis can offer rewarding insights, as it is hoped that the chapter will show.

Having offered a disclaimer first, what we will do in this chapter is introduce the reader to the approaches, models, and concepts of general equilibrium theory. We will need to use general-equilibrium (GE) analyses in subsequent chapters, and we want the reader to be familiar with the terminology, have a solid idea of what additional information can be expected from the GE-analysis, and understand why it is important to use GE analysis in some situations and less crucial in others. We will proceed in this section with the factual basis of general equilibrium problems, then offer a quick tour of the types of GE models available, and conclude with an equally brief tour of the types of questions GE analysis addresses.

5.1.1 *The facts*

The first major fact that can lead us to think in general-equilibrium terms is the interconnect-edness of markets. In the partial-equilibrium analysis we have seen the prevalence of opportunities for cross-price effects (although admittedly we have not looked empirically into the magnitude of those effects). If we study some phenomenon that affects a sufficiently large number of markets at the same time, the accumulation of the cross-effects could be important. This will be the case with taxation in particular. If a market we study is sufficiently large, events in it may be felt in other markets in ways that feed back on the market in question. In particular, when we study product markets, the effects may work on factor markets, causing the input prices facing our product market to change. Conversely, when studying a factor market, product prices may change and cause feedbacks on factor prices.

The second major fact that brings up a GE approach for consideration is the finiteness of

factor supplies. In partial-equilibrium analyses we use both demand and supply concepts, the latter describing for us how increasing demand for a good or a factor affects the cost terms on which we can expect to obtain it. The constraining factor behind rising supply curves is always input availability – factor supplies. As long as the factor supplies are unconstrained (their supply curves are flat), there’s little reason for product supply curves to slope upward in the long run. Supply curves by themselves, however, generally capture the bidding of various industries for factors; they do not contain the information, all by themselves, that at any particular time, there’s only so much, say, labor to be had in the economy, no matter what someone offers to get it. The same goes for capital and land. When we study an entire economy, we have to take account of these ultimately fixed factor supplies, and this affects the structure of models as well as results.

The third principal fact is that many of the economic systems we study have their manmade boundaries, and those territories define factor supplies.¹ Sometimes we will want to study how an entire economy, or a major portion of it such as its agricultural sector, responds to either some external change or to an internal policy. External economic policies have the national boundaries as their natural delimitation, and foreign trade policies frequently call for GE analysis. The territorial boundary of the ancient state, even if we might have difficulty drawing it on a contemporary map, implicitly brings with it fixed factor supplies.

5.1.2 *The models*

The question to keep in mind as we review the array of general equilibrium models available for use is “How should we slice reality to study it?” All models are simplifications and abstractions, and models sacrifice details dear to some scholars’ interests to obtain insights into more pervasive forces in a society. To offer a benchmark, let’s characterize a partial equilibrium model. The subject of a partial equilibrium model might be a firm. The variables whose behavior is to be explained might be the firm’s output of

one or several products and use of several of its inputs. The variables that the model accepts as given might be the technology, the prices of inputs and outputs (but not for a monopolist or monopsonist – or even for an oligopolist), some details about other firms, possibly something about the consumers of the output.

General equilibrium models sometimes focus on what are called sectors, and some of the GE models can be distinguished usefully along the lines of the number of sectors they use. Most of these are production sectors, frequently corresponding to industries or groups of industries. Nevertheless, if government is specified as an economic agent, it would be called the government sector. There might be a foreign “sector” in GE models involving but not focusing on foreign trade. But having identified what the term “sector” refers to, let’s begin with the simplest model of general equilibrium, the one-sector model. This model has a single production sector that produces an undifferentiated product, and typically it uses two factors, labor and either capital or land, in fixed quantities. With only one sector, there is no need for specification of demand conditions: the entire output gets consumed, and the incomes of land and labor are determined by available supplies of each and technical conditions in production. The model is useful for exposing the influences of the factor input ratio and the elasticity of substitution on the distribution of income across factors.

Two-sector models of general equilibrium have long been popular vehicles for studying a number of relationships: income distribution again, interindustry differences in technologies, interactions between product and factor prices, effects of increases in the differential growth of factor supplies, and others. Two production sectors are specified, and commonly two factors used in both industries. The production sectors are differentiated by their technologies, which may be parameterized (have their technology parameters set to particular values) to represent, for instance, industry and agriculture, importable and exportable goods, taxed and untaxed sectors, and so on, according to the problem of interest. Sometimes the model has been examined with

a factor unique to one of the industries. The demand side of the economy may be modeled explicitly, or product prices may be fixed exogenously on the grounds that the economy “takes” all its prices as determined in international markets. The model has had numerous applications in the analysis of issues in public finance (taxation) and international trade.

Possibly the most influential model of general equilibrium is the framework of Leon Walras. It has been a major vehicle for purely theoretical analysis but has supplied the basic format for GE models developed for more applied purposes – specific questions about tax policy or international trade. The structure of the Walrasian model can be viewed from two perspectives: that of a number of individuals buying and selling in different markets, and that of the markets themselves, in which demands and supplies of various goods are equilibrated by movements in prices. The Walrasian tradition in GE modeling has been adapted to examine only some components of allocation problems at a time. In the more abstract study of GE, some models have contained only exchange, others only production, and a few both consumption (exchange) and production.

Multiple-sector general equilibrium models, other than the abstractions of the basic Walrasian model, have been constructed with empirical uses in mind. The earliest of these was the input-output (i-o) model of Wassily Leontief, which specified fixed input coefficients for multiple production sectors and let many sectors produce intermediate products (inputs to other productive sectors) as well as final consumption goods. Factor supplies are fixed, and a set of demands for final consumption goods by households and government provide a “pull” for the flow of inputs and outputs through the model. The development of scientifically based national accounts beginning in the 1930s permitted the construction of empirical input-output tables (tables with ratios of various inputs to a unit of output for each production sector) of national, and eventually even regional and metropolitan economies. These models were solvable with matrix algebra computations, which was a major attribute in their favor in the early

days of automated computing. They have been useful for projecting the small-area employment impacts of changes in final demand.

Computable general equilibrium models (CGEMs), which use flexible forms of production functions such as the Cobb–Douglas, constant elasticity of substitution (CES), and eventually even the translog, have begun to supplement the i-o model, although the i-o approach is still popular. The construction of complete systems of consumer demands in these models proved at least as significant a challenge as the production side. These models also are solved numerically, but with more sophisticated simultaneous nonlinear equation algorithms, and their production and consumption parameters are usually calibrated to correspond to those of an actual national economy at a particular time. These models have been used to study tax policy, urban development in developing countries, trade policies, various questions of urban and agricultural development, and so forth. They are quite applied, in the sense that they are developed to study specific questions, generally of policy or of historical development.

5.1.3 *The questions*

We have noted the kinds of questions each of the types of GE model mentioned above addresses, but some recapitulation will be useful. The highly abstract GE models have addressed characteristics of the equilibrium of specific models of an economy: existence, stability, and uniqueness. These questions do not refer to whether the economy of, say, the United States or France had an equilibrium in 1994 and whether it was stable. Those are far different issues. General equilibrium models can include or exclude many features, and those questions are frequently of interest regarding particular features. For example, externalities and public goods (both to be discussed in the next chapter), fixed costs in production, increasing or decreasing returns to scale in production all cause difficulties for simple, marginal-cost pricing; we know that “real” economies are rife with those features, we know that they have been

for centuries, and we know that the economies have continued to “chug along” without either imploding or exploding (even including in our observation set the bouts of both hyperinflation and depression of the late nineteenth, twentieth and early twenty-first centuries) – so it seems reasonable to infer some measure of stability in the real world. If inclusion of some of the logically more-difficult-to-handle features of reality cause the equilibria of the models we use to study these real-world economies to not exist or to be unstable (which would give them multiple, rather than unique, equilibria), our models have some problems – not necessarily the world. But since the models demonstrate their usefulness by helping us understand the observational world, they should be able to incorporate these “irregular” but common features of that world without falling apart. In this sense, part of general equilibrium theory is about models rather than the observational world. We will illustrate some of these issues below in a very simple version of the Walrasian model of general equilibrium.

There may appear to be some coincidence of interests between the questions of existence and stability and those involved in the issue of the “collapse of complex societies” paradigm introduced recently into the study of ancient civilizations in the Mediterranean-Aegean area and elsewhere (Tainter 1988). In fact, however, the time scales of GE models, and their behavior in the event of nonexistence or instability of equilibrium, differ substantially from the events implied in the archaeological records of civilizations such as the Minoan and Mycenaean that have been transformed and possibly destroyed and forgotten over a period of a century or so. General equilibrium models might prove quite useful in examining the various hypotheses regarding the dissolution (if that really was the case) of these societies, but the issues of existence and stability of equilibrium in abstract GE models are several steps removed from these empirical problems.

The one-sector and two-sector GE models have proven useful for analyses that are still relatively abstract but less so than the existence, stability,

and uniqueness issues (although they can be examined for those properties). They have yielded insights about the economy-wide determinants of the functional distribution of income, the relationships between interindustry technological differences and the behavior of factor and product prices, and in general the indirect effects of economic changes that work through markets other than those immediately affected. As the numbers of goods and factors have increased, it has proven increasingly difficult to get unambiguous results out of these models, but their basic insights into two-sector/two-factor circumstances have shown how the allocation mechanisms operate in the more intricate situations.

The final set of questions we identify are what we call the more applied ones: the impacts of tax policies, the effects of international trade on a single country's resource allocations, the effects of economic growth on sectoral allocations of factors and of various events on economic growth. The one-sector and two-sector models have yielded useful general answers in these topics, but for more specific answers, multisector, numerically solved models have been required.

Before proceeding to a specific general equilibrium model, the issue of disequilibrium should be addressed. Many periods in antiquity are considered to have been times of quite generalized disequilibrium for the societies under study: the various Intermediate Periods in Egypt, LH IIIC in the Greek area, and so on. Why not use models of how people behave economically in periods of disequilibrium? Frankly, because models of choice made in disequilibrium situations – such as when an economic agent purchases an amount of a good that differs from the amount that his or her demand functions would say to purchase at a particular transaction price – involve difficult mathematics, are not particularly tractable, and have not been made particularly accessible yet. The subject is nonetheless important. The solutions of both static and dynamic (changing over time) economic models are equilibria satisfying all the relevant tangencies described in previous chapters, but how economic agents move from one equilibrium to another when

conditions change is not addressed in these models. The issues of concern are whether movement from one equilibrium position to another is stable – that is, will agents actually reach a new equilibrium when the original one is displaced – and how long it takes – that is, will a new equilibrium be reached before a new shock is likely to appear? The answer to this indicates whether an economy will spend most of its time close to equilibrium or in disequilibrium. An important aspect of disequilibrium is that agents' expectations are incorrect. Arbitrage is typically left as a black box that economists assume moves disparate prices toward one another, but the choices involved in arbitrage have not been modeled widely. It would be tidy to be able to summarize the modeling to date on this topic, but it must remain a subject left untreated explicitly here.²

5.2 The Walrasian Model³

We will start with a very simple economy – just two people involved in exchange of two goods. We can think of the goods as just collected where the consumers stand or as having been “dropped on them” from outside. (This is part of the abstractness of many GE models; they can be extended to include production, but a number of important aspects of exchange can be studied by holding production constant for the time being.) Both consumers begin our analysis with an endowment of both of the goods, $\bar{x}_i^j$, which is read as individual j 's endowment of good i ; we can identify the two individuals as A and B and the two goods as 1 and 2. Now we introduce the concept of excess demand, a construct much used in GE analysis. It is the difference between an individual's desired consumption level of a good and his endowment of it. If we work with a multi-period model, the endowment at the beginning of each period is just what the individual had left over at the end of the previous period, plus any new income at the beginning of the period, measured in terms of these goods rather than money. We can define excess demand

as $E_i^j \equiv x_i^j - \bar{x}_i^j$, in which x_i^j is the desired level. If E_i^j is positive, individual j has a demand for good i , and if $E_i^j < 0$, she is a supplier of good i . When $E_i^j = 0$, individual j 's demand for good i equals her supply of it, and she is said to be in equilibrium.

Our individual consumer's budget can be expressed in terms of excess demands because we know that her initial endowment defines the amount of consumption she can attain by exchanging some of one of her goods for some of the other at some price ratio. Let's begin by putting her desired consumption levels on one side of a "balance sheet" and her endowments on the other: $p_1 x_1^j + p_2 x_2^j = p_1 \bar{x}_1^j + p_2 \bar{x}_2^j$, but note that this really doesn't tell us that she consumes in exactly the same pattern as her endowment. We can rearrange this expression in terms of excess demands to get the information that, when she reaches an equilibrium, the sum of her excess demands is zero: $p_1(x_1^j - \bar{x}_1^j) + p_2(x_2^j - \bar{x}_2^j) = 0$. Both individuals A and B have similar budget-consumption relationships, and if we add them together, we get $p_1 \sum_j E_1^j + p_2 \sum_j E_2^j = 0$, or equivalently, $p_1 E_1 + p_2 E_2 = 0$, which says that the sum of the value of the excess demands of each good in the market is zero.⁴ This is known as Walras' Law. One of the implications of this relationship is that if all markets but one are in balance, the last one will be in balance as well. So, if there are n markets, we need to know about equilibrium conditions in $n-1$ of them.

It may be useful to extend this demonstration to three goods and just stay with superscripting for the identity of individuals. Each individual has a utility function $u^j(x_1^j, x_2^j, x_3^j)$, which he or she maximizes subject to a budget constraint, which we express as the sum of excess demands for the three goods: $p_1(x_1^j - \bar{x}_1^j) + p_2(x_2^j - \bar{x}_2^j) + p_3(x_3^j - \bar{x}_3^j) = 0$. Notice that we have done one of two equivalent things regarding the prices of the goods: we have either measured p_1 and p_2 in terms of p_3 , or set p_3 equal to one, which amounts to the same thing; at any rate, the price of the third good is the numeraire. Then using this information, we can

find each individual's demand functions for each of the goods: $x_1^j = d_1^j(p_1, p_2)$, $x_2^j = d_2^j(p_1, p_2)$ and $x_3^j = d_3^j(p_1, p_2)$. The total demand for each good is the sum of the demands of each of the individuals j , and that must equal the total supply of each good: $\sum_j d_1^j(p_1, p_2) = \bar{x}_1$, $\sum_j d_2^j(p_1, p_2) = \bar{x}_2$, and $\sum_j d_3^j(p_1, p_2) = \bar{x}_3$. The supplies of the three goods are given parametrically (they're exogenous – by excluding the production side of the problem, we just assigned values to them; the model as it stands doesn't have to determine those values). The model adjusts the prices p_1 and p_2 to equilibrate the sum of the individual demands to the total amounts available of goods 1 and 2, $\bar{x}_1$ and $\bar{x}_2$; we know that the third market will "clear" – that is, the quantity demanded of it will equal the supplies available. What the model determines is the prices of goods 1 and 2, in terms of the price of good 3, which we've set equal to one. If we substituted those equilibrium prices back into the individuals' demand functions for goods 1 and 2, we could find the allocations of goods 1 and 2 to each individual, and residually each individual's consumption of good 3.

Figure 5.1 shows the issue of existence of equilibrium graphically, for the case of two goods. We measure the excess demand for good 1 positively by moving to the right on the horizontal axis. Excess demands to the left of the price axis (the zero point on the excess demand axis) are negative and correspond to excess supplies of good 1. The diagonal line going from southeast to northwest is the combination of relative prices p_1/p_2 associated with particular levels of excess demand. They have the relationship $\Delta(p_1/p_2) = f(E_1)$, which says that the change in the relative price p_1/p_2 is a function of the level of excess demand for good 1. Figure 5.1 shows that if E_1 is positive (>0), $\Delta(p_1/p_2)$ is also positive: the relative price of good 1 rises when there is an excess demand for good 1. If E_1 is negative (<0), $\Delta(p_1/p_2)$ is negative: the relative price of good 1 falls when there is an excess supply of it. The demands for good 1 (and by implication for good 2) as Figure 5.1 is drawn yield a unique equilibrium at p^* , and the equilibrium is stable: try moving excess demand away from the equilibrium point at p^* a bit, and

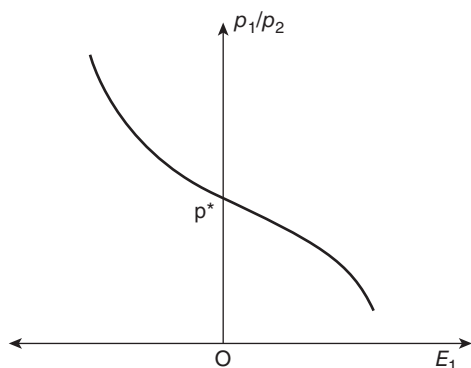


Figure 5.1 Excess demand, unique stable equilibrium.

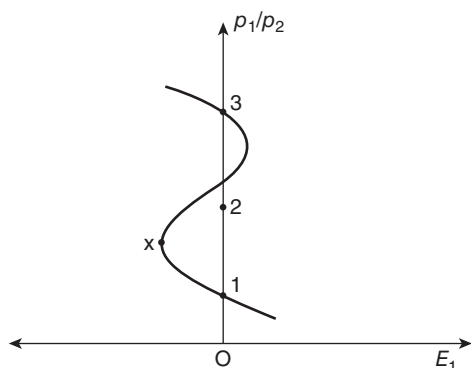


Figure 5.2 Excess demand, multiple equilibria.

the slope of the line shows that it has a tendency to move back toward the equilibrium point from which it was disturbed.

Not so in Figure 5.2, which shows the case of multiple equilibria, two of which are stable and one of which is unstable. At a value of $p_1/p_2 = 1$, there is an equilibrium of supply and demand. If we were to perturb the price ratio upwards to the level indicated by x along the excess demand curve (possibly representing a shift in preferences of consumers toward good 1), the relative price would return to the value 1, but if preferences shifted far enough toward good 1, the shape of the excess demand curve would propel the price ratio away from the stable equilibrium at

$\Delta(p_1/p_2)$ and up to the unstable equilibrium at about $p_1/p_2 = 2.5$. From the slope of the excess demand curve at that point, and our knowledge that $\Delta(p_1/p_2)$ increases when E_1 is positive, the price ratio will keep increasing until it reaches a new, stable equilibrium at about the value of $p_1/p_2 = 3$. If p_1/p_2 moves above that point, it will have a tendency to return to it since above that point E_1 is negative and $\Delta(p_1/p_2)$ will accordingly be negative. What can give an economic system such as this one a tendency to unstable and multiple equilibria? Substantially differing marginal propensities to spend on the goods (income effects), between individual consumers in this two-consumer example, or between major categories of individual, can suffice. When the relative price of, say good 1 rises, the substitution effect pushes both individuals in our two-person case away from good 1. For the consumer who was a net demander of this good, the income effect is negative, but for the one who was a net supplier, it is positive. The actual sizes of the gain and loss between the two individuals are about the same, and if their marginal propensities to spend incremental income on good 1 are about the same, the effects will cancel. If they are substantially different, *and* the difference itself is large, we can get multiple equilibria.

This demonstration also shows what is meant by stability or instability of an equilibrium. Stability means that, if the equilibrium values of the variables characterizing the equilibrium (the "endogenous" variables) are perturbed, they will tend to move back toward the original equilibrium. To determine how quickly they would return we would have to add specifications that characterized the pace of the movements. Instability means that the same sort of perturbation would send the values of the endogenous variables shooting through another, unstable equilibrium to a different, stable equilibrium. Instability does not mean that the system would either collapse on itself or explode, although the nearest stable equilibria need not be particularly close to each other. Instability implies multiple equilibria, two of which are stable for every one that is unstable; and multiple equilibria imply an unstable equilibrium, one for every two stable equilibria.

If we wanted to characterize, say, the destruction of the Mycenaean palatial “system” (whether it was a unified kingdom or separate kingdoms, a particularly unified economic system regardless of its political structure), we would have to specify a number of economic relationships for those societies (or “that society”). Those specifications would have to include some variable or variables whose values could “jump” and would correspond to some of the things that we think “went south” after about 1200 B.C.E. – income comes to mind naturally as such a variable, although that might need to be characterized as a combination of prices and quantity endowments. Or the changes the Mycenaean civilization experienced might not involve any unstable-equilibrium issues at all. That concept could be a red herring.

5.3 Exchange

The reader will have noticed that in the abbreviated presentation of the Walrasian model of pure exchange (we left out production) there was no sense of “surplus,” or people offering “what they didn’t need” themselves for other things they did “need.” (“Need” is not defined in economics; the closest that word can be approximated with an economic concept is with “demand,” used as a noun. Since we all get by without things we need, what does “need” mean?) People began with endowments and preferences and engaged in exchange until demands equaled supplies. This is a critical difference between the economic analysis of exchange and trade and the notions of “surplus exchanges” that appear commonly in the study of ancient civilizations.

We will show a graphical presentation of these barter exchanges, using the Edgeworth box introduced at the end of Chapter 2. In the present application, in Figure 5.3, we are measuring goods on the axes of the joined graphs (we used factor inputs in the previous application). The point labeled R, toward the upper left corner of the box, represents the initial endowment of our original two individuals with goods x_1 and x_2 . Individual A’s indifference curves for goods

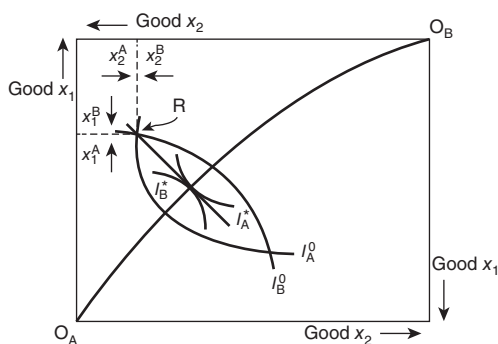


Figure 5.3 Analysis of two-person/two-good exchange with an Edgeworth box diagram.

x_1 and x_2 are measured outward from origin O_A , and those of individual B are measured downward from origin O_B . The contract line, which need not be monotonic but does not cross the diagonal line between the two origins, is the locus of tangencies between the families of indifference curves of A and B. A straight line through both the endowment point, denoted by the point at R, and a tangency of the two indifference curves (there is only one such line) represents the relative price of goods x_1 and x_2 , which lets A and B exchange parts of their initial endowments in a pattern that will yield each the highest utility he can reach without reducing the utility of the other. At the endowment point, the two individuals’ indifference curves cross, so that allocation can be improved upon.

Each point along the contract line indicates a combination of x_1 and x_2 for A and B that gives them the same marginal rate of substitution between the two goods, but given their initial endowments, the marginal rate of transformation that takes the initial endowments to the allocation that equalizes MRS for the two consumers is not equal to the MRS. For an efficient allocation, $MRT = MRS$, and there is only one allocation that yields this joint equality, the tangency location of indifference curves I_A^* and I_B^* . You can see, however, that if we found the price line with the slope of this $MRT = MRS$ condition, a reallocation of the initial endowments would permit an efficient exchange allocation at any

point on the contract line. This is the efficiency basis of many taxation-and-redistribution plans that we will introduce in the next chapter, on public economics.⁵

5.4 The Two-Sector Model

We will show some of the basics of the two-sector model of production and income distribution in this section.⁶ We have already introduced one of the vehicles used to convey it – the Edgeworth–Bowley box (frequently called simply the Edgeworth box). Another graphical device we will use is the Lerner–Pearce diagram, which is a more intricate version of a production isoquant diagram adapted to two production processes at a time instead of just one.

5.4.1 The basics with the Lerner–Pearce diagram

In general equilibrium there is a unique relationship between commodity prices, factor prices, the pattern of production, and the distribution of income. The Lerner–Pearce diagram uses the fact that constant returns to scale in production let a single isoquant satisfactorily represent the entire expansion path of the production function. The ratios of marginal products of each pair of inputs depend only on the ratios of the inputs. In Figure 5.4, a series of isoquants along either

of the two expansion rays, OR_X or OR_Y , would be tangent to parallel factor price lines (we only show one of the factor price lines, to keep down the clutter in the figure). Figure 5.4 shows the isoquants of two industries, X and Y, each of which uses two factors, land, L , and labor, N . Industry X is relatively land intensive. In drawing the isoquants tangent to a common factor price line, we are implicitly defining the unit cost of each to be the same, and equal to distance OM measured in terms of land, or OM' in terms of labor; we can define the units so that these isoquants define one unit of the output of each industry. Point E along the factor price line MM' is the economy's endowment of land and labor. Lines OR_X and OR_Y are the factor price rays – lines from the origin of the graph showing the factor input ratio. (The Lerner–Pearce diagram can get pretty cluttered, so we add components a few at a time in separate diagrams so the reader can find the new elements more easily and see what they are doing more clearly.)

Figure 5.5 shows the effect of a change in the relative price of land and labor. To raise the relative price of land, we can rotate the factor price lines separately around each isoquant (by virtue of the representativeness of any isoquant). For the X industry, we have factor price line M_X and for the Y industry, M_Y . The factor input rays OR_X and OR_Y both move in a clockwise direction to OR'_X and OR'_Y . The production costs of the two

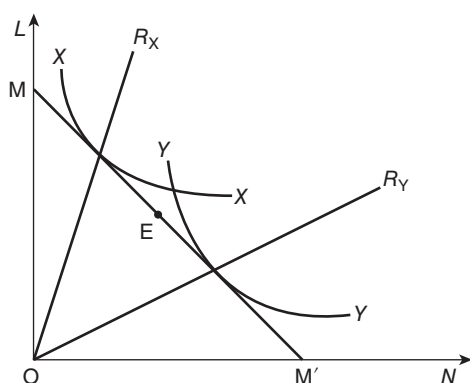


Figure 5.4 The Lerner–Pearce diagram.

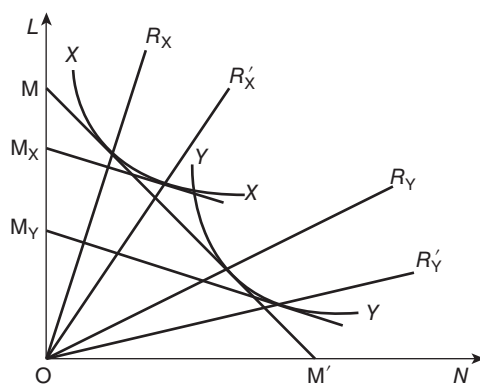


Figure 5.5 Change in factor prices in the Lerner–Pearce diagram.

goods now differ: OM_X for a unit of the X good and OM_Y for a unit of Y; so the relative price of the land-intensive good increases when the price of land increases. More generally, there is a one-to-one relationship between the relative costs (prices) of goods and the relative prices of factors used in producing them. An increase in the relative marginal productivity (price) of a factor produces a rise in the relative price of the good that uses it intensively. A fixed factor endowment puts limits on the range of commodity and factor price ratios that an economy can reach and still maintain full employment of both factors. (Perhaps this is a good time to redefine the term “intensity.” “Factor intensity” can be measured as the physical ratio of one input to another. Comparison of factor intensities across activities – industries – is meaningful when the factor-price ratios are the same across industries; in that circumstance, we compare the cost-minimizing ratio of factors across industries. The factor intensity concept only becomes useful when we compare it across activities.)

Let’s examine the information that the endowment point, E, lets us derive from the diagram. The line OR through E is the economy-wide factor-endowment ray. The factor utilization rays for the two industries, OR_X and OR_Y , have to fall on opposite sides of this ray. Rather than include the isoquants for the X and Y industries, in Figure 5.6 we have just identified their tangency

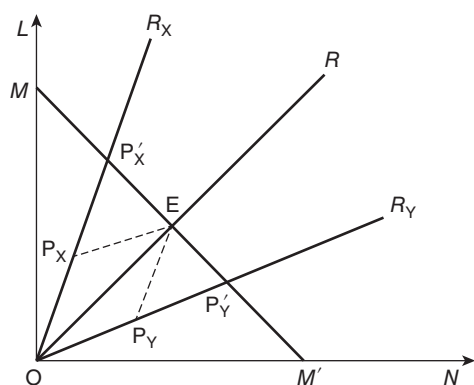


Figure 5.6 Relationship between commodity prices, factor prices, and the pattern of production.

points as P'_X and P'_Y . With the information given by the factor endowment, we can determine (by vector addition) the unique relationship between commodity prices, factor prices, and the pattern of production – the allocation of factors to industries. From point E, we draw a dotted line parallel to ray OR_Y , back to its intersection with the factor utilization ray for the X industry, OR_X , at point P_X . Similarly, draw a line from E parallel to ray OR_X , back to an intersection with ray OR_Y at P_Y . The distance OP'_X along ray OR_X is a measure of “national” income in terms of good X. The ratio OP'_X/OP'_X is industry X’s share of national income, measured in terms of good X; $P_X P'_X/OP'_X$ is the share of industry Y in national income measured in terms of good X. Along ray OR_Y , measured in terms of good Y, the value of national income is OP'_Y , the share of industry Y is OP'_Y/OP'_Y , and the share of the X industry is $P_Y P'_Y/OP'_Y$.

In a relationship we do not show graphically, but which you can envision from comparison of Figures 5.5 and 5.6 – or draw for yourself – as the relative price of land rises (MM' twists counterclockwise around point E), P_X shifts to the northeast and P_Y to the southwest, which implies an increase in the production of X and a decrease in Y production. This in turn implies a higher relative cost (price) of good X; so increased production of X causes a rise in the relative cost of production in terms of the amount of good Y foregone. This is the cost relationship that causes a transformation curve between goods X and Y to “bow out” away from the origin (see Figures 2.17 and 2.19).

So far, we have worked only with the production side of this economy. We have been able to see how different factor price ratios would produce different product price ratios and different factor allocations, but we need information on demand conditions to determine just which particular production and distribution pattern will emerge. When we have worked with budget constraints, they have been defined in terms of commodities at a given set of commodity prices. However, that budget constraint is implicitly a constraint defined in terms of factors at the factor price ratio corresponding to the commodity

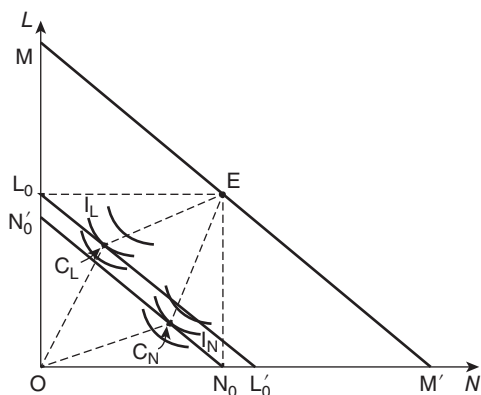


Figure 5.7 Correspondence of commodity and factor prices.

price ratio. We can decompose goods demanded, subject to the commodity / commodity price constraint, into the factors required to produce them by making use of the optimal factor utilization ratios corresponding to the factor price ratio. Preferences defined over commodities are equivalent to a set of preferences defined over factors. We can see this in Figure 5.7, which uses indifference curves between land and labor that are equivalent to the direct indifference curves between goods X and Y. From the endowment point, E, draw a line to the land axis, marking off the land endowment at L_0 , and make the corresponding projection of the labor component of point E to the labor axis at point N_0 . From L_0 on the land axis, draw a land-owners' budget line parallel to the aggregate factor price ratio, MM' . That budget line is $L_0K'_0$. The corresponding budget line for labor owners is N'_0NL_0 . Now, let's put a pair of indifference curve families in the diagram, one for land owners (spending income from land) and one for labor (suppliers of labor). We identify the tangency points as C_L and C_N for land and labor. Next, draw lines from C_L and C_N to the origin; then, from point C_L on the land budget line, draw a line parallel to OC_N and another from point C_N on the labor budget line parallel to OC_L . As the indifference curves are drawn in Figure 5.7, they fortuitously add up,

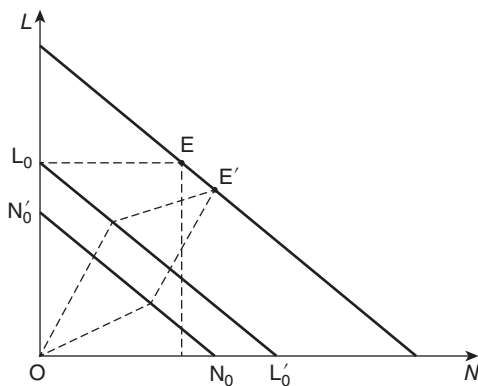


Figure 5.8 Excess demands and supplies in the Lerner-Pearce diagram.

by vector addition, to the endowment point, E. They need not, however, as Figure 5.8 shows. In that figure, E is the economy-wide endowment and E' is the endowment implied by the factor demands emanating from the current factor price ratio. There is an excess demand for labor and an excess supply of land at E' , which in turn imply an excess demand for the labor-intensive commodity, Y, and an excess supply of the land-intensive commodity, X. With the economy's current factor endowment, the relative price of labor has to rise until E' and E coincide, as they did in Figure 5.7.

5.4.2 Growth in factor supplies

We use the Lerner-Pearce diagram to explore the mechanics of adjustment induced by changes in factor supplies. Differential growth in factor supplies is of particular interest, so we can augment just one factor, say labor, and observe the economic adjustments. To simplify the view, we assume that commodity prices are fixed. Consequently, factor prices will be fixed, and the optimal factor-utilization ratio will be unchanged. (We'll relax some of these assumptions below.) In Figure 5.9, the increase in the labor force moves the endowment point from E to E' . OR_X and OR_Y remain the same, but the changed location

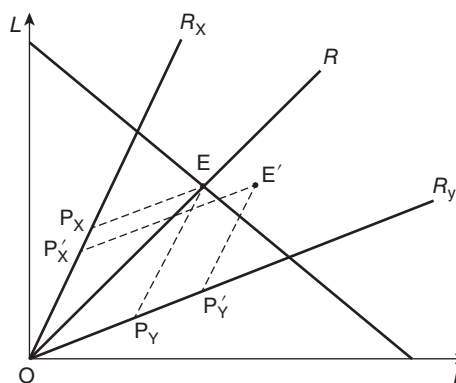


Figure 5.9 Growth in factor supplies in the Lerner-Pearce diagram.

of E' puts the projections of the parallels to the factor-utilization rays farther out along OR_Y and closer to the origin along OR_X : out to P'_Y for good Y and back to P'_X for good X. With the constant factor utilization ratios, the increase in labor is absorbed into both industries, but in both industries, additional land is required to accompany the labor in the production process. With the fixed factor utilization ratios (following from the fixed factor price ratio which itself followed from the fixed commodity price ratio), only one of the industries can expand: the land required to accompany the new labor has to come from somewhere. To maintain the marginal rate of substitution between land and labor at the constant factor price ratio, additions to labor used in the Y industry must be accompanied by sufficient units of land, coming from the X industry, but releasing land from the X industry also releases the labor that was used with it, which also must go to the Y industry. To absorb these units of land and labor, the Y industry must expand absolutely and the X industry must contract absolutely. Alternatively, if we tried to fit the new labor into the X industry, the decreased production in the Y industry would not release enough land for each unit of labor it released to maintain full employment of labor. Consequently, the Y industry expands absolutely and the X industry retrenches

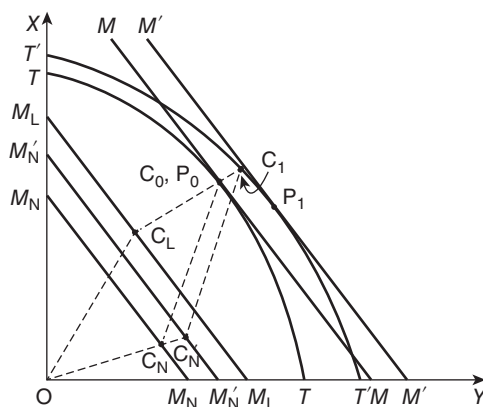


Figure 5.10 Price changes following growth in factor supplies.

absolutely. In general, the industry that uses intensively the factor whose supply has expanded will expand absolutely and the industry that uses the other factor intensively will contract absolutely. This is known as the Rybczynski theorem (Rybczynski 1955).

If the original configuration of outputs was fully consumed by this economy, the new configuration will be unsustainable at the old commodity and factor prices. We can show the demand adjustments with a transformation-curve diagram. In Figure 5.10, point C_0, P_0 on the transformation curve TT represents the initial combination of X and Y produced at the commodity price ratio represented by NN' . Land's budget line is $M_L M_L$ and labor's is $M_N M_N$; points C_L and C_N represent the tangency points of land and labor indifference curves (indifference between goods X and Y rather than between factors in this diagram) and their budget lines. Vector addition gives a consumption and production point on transformation curve TT at point C_0, P_0 , which is the point of tangency with the commodity price line MM (associated with the factor price line MM in the Lerner-Pearce diagram of Figure 5.7).

The increase in labor supply extends the new transformation curve $T'T'$ more in the

direction of good Y than good X, since Y is the labor-intensive good. At the original factor prices, the new production point is P_1 on $T'T'$. The increment in income is all income to labor since factor prices have remained constant, so the labor budget line shifts out to $M'_N M'_N$ while land's budget line remains unchanged at $M_L M_L$. Labor's consumption moves out to C'_N while land's remains constant at C_L . By vector addition, the new, combined, desired consumption point is C_1 on the new transformation curve $T'T'$. If we were to draw a combined indifference curve (combined for capital and labor), it would not be tangent to the new transformation curve at C_1 , but somewhere between C_1 and the new production point, P_1 . Clearly, commodity prices will have to adjust: the relative price of good Y, which was expanded by the labor supply increase, must fall, twisting $M'M'$ counterclockwise. That will alter factor prices, which will change the intercepts of the budget lines of land and labor, and through several iterations, bring C_1 and P_1 into coincidence, with lower relative prices of both good X and labor.

5.4.3 Technical change

There is considerable scope for technical change in the two-sector model. Each industry could undergo neutral, land-saving (= labor-using), or labor-saving (= land-using) technical change. Neutral technical change leaves the optimal factor utilization ratio unchanged at constant factor prices. Land-saving technical progress reduces the land-labor ratio at constant factor prices, and labor-saving technical change does just the opposite. We will give a brief, graphical treatment of two cases, neutral and labor-saving technical change, both in the land-intensive industry, X.

Figure 5.11 shows neutral technical change in the X industry. The initial X-industry isoquant is shown by XX ; the technical change reduces the quantities of land and labor required to produce a single unit of X, which is shown by the shift back toward the origin of the new isoquant, $X'X'$.

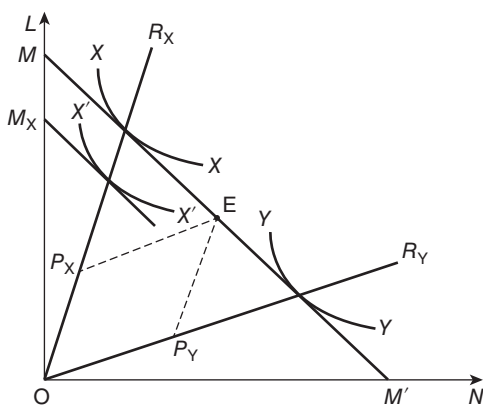


Figure 5.11 Neutral technical change.

Factor price line M_X is parallel to the original factor price line; this is by our own assumption, as demand conditions might cause it to change, but to show the adjustments step by step, we let it stay constant for now. The unit cost (price of good X) falls from OM to OM_X . If factor prices are to remain constant, there must be no reallocation of factors between the X and Y industries; that is, the X industry must be able to expand its output enough to absorb all the units of land and labor it originally employed in producing the smaller output with the more expensive technology. Whether this will occur or not depends on demand conditions and the fate of commodity prices. To keep commodity prices unchanged the demand for X must expand proportionally to the change in output and the demand for Y must remain constant. An elasticity of demand for X equal to 1 (including both income and substitution effects) will yield this constancy; a value greater than 1 will yield an excess demand for X and an excess supply of Y, which will produce a rise in the relative price of X and thence in the relative price (per period rental rate) of land. A unit demand elasticity would also keep the relative and absolute income share of land constant. Vice versa if that demand elasticity is less than 1; an inelastic demand would reduce the price of good X, the price (rental rate) of

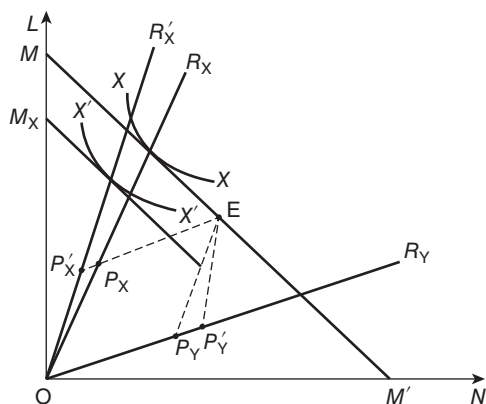


Figure 5.12 Labor-saving technical change in the land-intensive industry.

land, and the relative and absolute share of land in “national” income. Thus the consequences of a change in technology can be broken rather cleanly into pure production effects and pure demand effects.

For labor-saving technical progress in the land-intensive industry, we turn to Figure 5.12. The technical change – the switch to the $X'X'$ isoquant – pushes the factor utilization ray toward greater land-intensity, from OR_X to OR'_X . If factor utilization in the X industry remained the same as before the technical change, isoquant $X'X'$ would cut the commodity price line M_X at the intersection with ray OR_X , which would leave the marginal rate of transformation in production lower than the marginal rate of substitution in the market for factors. At the old factor utilization ratio, the price (rental rate) of land would have had to rise; since we’re keeping the factor price ratio constant for a while by assumption, the factor utilization ratio has to change to keep $MRT = MRS$. The production point in the Y industry moves out from P_Y to P'_Y , which represents an expansion of the Y industry. Whether the X industry expands or contracts remains an open (empirical) question, hinging on both production and demand conditions. Whether the exit of inputs from the X industry is large

enough to outweigh the increased productivity of inputs depends on relative factor intensities (note that substitutability in production is not an issue yet because we’re holding factor prices constant; so this is a Rybczynski-like effect at work). The critical value of the demand elasticity for X for the movement of income shares is less than one in this case. That is, if that elasticity falls below some particular value less than one, the prices of good X and land will fall, and the share of land in income will fall as well.

5.5 Existence and Uniqueness of Equilibrium⁷

The existence of equilibrium is a logical matter, not an empirical one. It is a test of the logical consistency of the models and at its most basic of the entire notion of a competitive economy as the model has come down from Walras and has been used as a workhorse of theorizing in fields from international trade to monetary theory. It should be kept in mind, however, that simply because a general equilibrium model possesses a solution this doesn’t necessarily imply that it’s a particularly good model, but if it doesn’t have one, the model doesn’t warrant serious consideration. At an extremely simple level, people have tried counting equations and unknowns on the assumption that equality of the two means that the set of equations might have a solution.⁸ As many an eighth grader has discovered, not every set of simultaneous equations with equal numbers of variables and equations has a solution: satisfying one relationship may preclude satisfying one or more others. In a very general model purporting to characterize many sectors of an economy, counting equations and unknowns, and trying them out one by one, is entirely inadequate; it would involve empirical specification with inadequate knowledge and would not offer any theoretical insight into logical problems.

The principal method employed to study existence has been an advanced branch of mathematics known as convex analysis or topology,

with specific tools called fixed-point theorems. It would not pay dividends to go into any details of fixed point theorems here, other than to say that the fixed point is a vector of prices emerging from excess demand or excess supply functions, introduced in section 5.2. If such a fixed point can be shown to exist under certain conditions, it means that a vector of prices exists that will solve the full set of relationships characterizing consumption and production in the model of an economy. If a fixed point does not exist, it means that under the conditions included in the model, an equilibrium does not exist – some producers will want to supply more or less of some goods than consumers want. However, it is possible that an equilibrium cannot be said to exist under certain conditions such as with increasing returns in production, or consumer preferences that cannot be easily averaged.

It has long been established that Walrasian general equilibrium for a competitive economy and complete markets exists. This is a strong set of conditions and leaves many questions. First, the competitive part. The competitive stricture can be interpreted as a kind of latent competition as characterized by the contestable market concept, so it is not as strong a condition as it may first seem. Beyond that, existence has been proven with some monopolistic and monopsonistic agents in the economy, as long as the increasing returns to scale underlying the monopoly are not too strong. Under more general conditions of increasing returns in production, equilibrium can fail to exist, but exploration of the failure has shown that such an economy can come close to equilibrium, in the sense that the remaining excess supplies or excess demands are small relative to the overall size of the economy. Externalities are cases of missing markets, and equilibrium has been found to exist under conditions of missing markets. Similarly with uncertainty and multi-period models. In some cases, restrictions have had to be applied to production technology and consumer preferences to yield existence, and this type of examination continues to be an active research field.

In general, when a new form of general equilibrium model is introduced, such as with some of the computable general equilibrium models

introduced in the following section, the matter of existence has been explored to assess whether the model is logically consistent. Once existence has been proven for a type of model, however, further applications of that model or variants of it are not typically subjected to such examinations. Sometimes that involves an element of faith.

Uniqueness of equilibrium is considerably more demanding than simple existence. However, the closer to a perfect and complete set of markets specified in a model – implying no increasing returns to scale in production, uniqueness is more likely. Increasing returns and important externalities can result in nonuniqueness of equilibrium. In contemporary monetary models, with endogenous expectations, uniqueness cannot be guaranteed and in fact many equilibria are theoretically possible. Overall, a smaller number of possible equilibria is preferable to a larger number simply because the range of predictions from such a model is narrower and more informative.

What does this have to do with ancient economies? Scholars have proposed models of various ancient economies, from the temple economies of ancient Mesopotamia to the redistributive economies of Pharaonic Egypt, Minoan Crete and Mycenaean Greece. These models have not been formalized to the point of symbolic representation of the relationships comprising them, but should they be so developed in the future, examination of existence of an equilibrium could serve as a useful consistency check.

5.6 Computable General Equilibrium Models

The models presented in this chapter have been quite abstract, but their purpose is to uncover and convey basic principles. In the 1970s, concerned with the different answers that partial equilibrium analyses could yield, a number of economists harnessed various versions of general equilibrium equation structures, fleshed out with actual numbers, to the period's new computing power. The earliest applications were to problems of economic development in the countries of what was then called the Third World.⁹ They have

gone on to model tax policies, energy policies and effects of global climate change in industrial countries.

In their developing country applications these models have identified two locational sectors – rural and urban – and a number of production sectors, although sometimes as few as agriculture in the rural sector and industry in the urban sector. To account for a 1960s–1970s phenomenon characterizing many Third World cities, these models would include an urban informal sector alongside an urban industrial sector. The informal sector represented traditional production methods and provided an equilibration device for the labor market, the urban unemployment rate adjusting rather than the urban sector's industrial wage to equilibrate the labor market: when the expected wage in the urban sector (the industrial sector wage times the probability of finding employment in that sector) equaled the rural sector's wage, migration from the countryside to the city would cease. This was neither a necessary nor a particular central feature of CGE models, but was a feature that scholars of antiquity may find appealing to apply to some situations of earlier times and places. Care must be taken in applications to antiquity, however, to allow the urban wage to adjust, as skill differentials between rural and urban occupations would have been far less than in mid-twentieth century C.E. Third World cities.

Continuing with the structure of CGE models, they specify factor supplies (labor, land, capital), production functions, demand functions, and market clearing (equilibrating) conditions that supply equal demand in all markets (omitting one market, via Walras' Law). An external sector can be specified, allowing for foreign trade. Economic growth can be modeled by specifying investment and population growth. The "computable" part of the name of this type of model gives away the fact that these models yield numerical solutions. The quantities of factor supplies are specified, and functional forms for the production and demand functions are specified, requiring that values of output elasticities for production functions and price and income elasticities of demand be specified. With the factor supplies and production and demand parameters specified, solutions for

the optimal allocation of factors to production of the various goods, the production of those goods, and the consumption of them by various categories of demanders, along with the factor and product prices that guide these allocations, can be obtained numerically.

The models are simulated as follows. A solution vector of prices and quantities is calculated numerically for one configuration of exogenous values. This solution is contingent on a configuration of factor supplies, production and demand parameters, and if the country is considered small,¹⁰ the prices of internationally traded goods. Then the value of one of the parameters or exogenous variables is changed, and a new solution is calculated. The values of the endogenous variables in the two solutions are compared, typically in the form of an elasticity with respect to the exogenous variable or parameter changed. This strategy allows the modeler to simulate alternative scenarios of what might have happened and see how the economy would have reacted, an exercise some of the users have called counterfactual history.

If this all sounds too easy and good to be true, there certainly are difficulties, which the literature has by and large acknowledged. Most obviously, the parameter values – the output elasticities for the production functions, substitution elasticities if CES production functions are used, and the demand elasticities – chosen may differ from the values in the economy the model is intended to represent. In applications to twentieth-century C.E. developing countries, an extensive empirical literature is available for guidance. Beyond the empirical faithfulness of the parameters of the model – of which there are many typically – calibrating the combination of fixed factor supplies and proposed parameter values must respect the relationships among some of the sets of parameter values (the demand parameters have intricate restrictions on their values – combinations of them have to "add up," as indicated in Chapter 3). These challenges surmounted, a CGE model is ready for simulation.

By 2007, CGE modeling found its way to Imperial Rome as an application (Gerachty 2007). I do not intend to review that application here,

but several comments are warranted. First, the determination of the parameter values is, as can be expected, a challenge, and while ancient historians may differ on the values chosen, they may want to offer preferred values of which they are aware. Second, the author kept the model small enough to solve “by hand,” although exactly what that term means is not clarified, and used Microsoft Excel to conduct sensitivity analyses with parameter variations. Third, the model

demonstrates the ability of economic analysis to offer a coherent story of events of the period without precise knowledge of everything about the economy. Fourth and finally, the story the model offers, whether ultimately judged to have been more or less correct or more or less incorrect, is based on general equilibrium interactions that would be likely to go unobserved in studies of any single question.¹¹

References

- Alchian, Armen A., and William R. Allen. 1969. *Exchange and Production Theory in Use*. Belmont CA: Wadsworth.
- Aliprantis, Charalambos D., Donald J. Brown, and Owen Burkinshaw. 1990. *Existence and Optimality of Competitive Equilibria*. Berlin: Springer.
- Archibald, Zofia. 2001. “Part II. Geographies and Place. Regional Economies.” In *Hellenistic Economies*, edited by Zofia H. Archibald, John Davies, Vincent Gabrielsen, and G.J. Oliver. London: Routledge, pp. 133–136.
- Arrow, Kenneth J., and Frank H. Hahn. 1971. *General Competitive Analysis*. San Francisco CA: Holden-Day.
- Bénassy, Jean-Pascal. 1982. *The Economics of Market Disequilibrium*. New York: Academic Press.
- Borglin, Anders. 2004. *Economic Dynamics and General Equilibrium; Time and Uncertainty*. Berlin: Springer.
- Davies, John K. 2001. “Hellenistic Economies in the Post-Finley Era.” In *Hellenistic Economies*, edited by Zofia H. Archibald, John Davies, Vincent Gabrielsen, and G.J. Oliver. London: Routledge, pp. 11–62.
- Debreu, Gerard. 1982. “Existence of Competitive Equilibrium.” In *Handbook of Mathematical Economics, Volume 2*, edited by Kenneth J. Arrow and Michael D. Intriligator. Amsterdam: North-Holland, pp. 697–743.
- Fisher, Franklin M. 1983. *Disequilibrium Foundations of Equilibrium Economics*. Cambridge: Cambridge University Press.
- Fisher, Franklin M. 2003. “Disequilibrium and Stability.” In *General Equilibrium: Problems and Prospects*, edited by Fabio Petri and Frank Hahn. London: Routledge, pp. 74–94.
- Gerachty, Ryan M. 2007. “The Impact of Globalization in the Roman Empire, 200 BC – AD 100.” *Journal of Economic History* 67: 1036–1061.
- Johnson, Harry G. 1971. *The Two-Sector Model of General Equilibrium*. Chicago IL: Aldine-Atherton.
- Johnson, Harry G. 1973. *The Theory of Income Distribution*. London: Gray-Mills.
- Jones, Donald W. 2000. *External Relations of Early Iron Age Crete, 1100–600 B.C.*, Archaeological Institute of America Monographs New Series, Number 4. Philadelphia PA: The University Museum, University of Pennsylvania.
- Kelley, Allen C., and Jeffrey G. Williamson. 1974. *Lessons from Japanese Development: An Analytical Economic History*. Chicago IL: University of Chicago Press.
- Kelley, Allen C., and Jeffrey G. Williamson. 1984. *What Drives Third World City Growth? A Dynamic General Equilibrium Approach*. Princeton NJ: Princeton University Press.
- Kelley, Allen C., Jeffrey G. Williamson, and Russell J. Cheetham. 1972. *Dualistic Economic Development: Theory and History*. Chicago, IL: University of Chicago Press.
- Krauss, Melvyn B., and Harry G. Johnson. 1975. *General Equilibrium Analysis; A Micro-economic Text*. Chicago, IL: Aldine.
- Manning, Joseph G. 2003. *Land and Power in Ptolemaic Egypt; The Structure of Land Tenure*. Cambridge: Cambridge University Press.
- Mas-Colell, Andreu. 1985. *The Theory of General Economic Equilibrium; A Differentiable Approach*. New York: Cambridge University Press.
- Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.
- McKenzie, Lionel W. 2002. *Classical General Equilibrium Theory*. Cambridge, MA: MIT Press.
- Moore, James C. 2007. *General Equilibrium and Welfare Economics; An Introduction*. Heidelberg: Springer.
- Nikaido, Hukukane. 1968. *Convex Structures and Economic Theory*. New York: Academic Press.
- Quirk, James, and Rubin Saposnik. 1968. *Introduction to General Equilibrium Theory and Welfare Economics*. New York: McGraw-Hill.

- Radford, R.A. 1945. "The Economic Organization of a P.O.W. Camp." *Economica* 12: 189–201.
- Reger, Gary. 1994. *Regionalism and Change in the Economy of Independent Delos*, 314–167 B.C. Berkeley CA: University of California Press.
- Rybczynski, T.M. 1955. "Factor Endowments and Relative Commodity Prices." *Economica* 22: 336–341.
- Schultz, Theodore W. 1975. "The Value of the Ability to Deal with Disequilibria." *Journal of Economic Literature* 13: 827–846.
- Tainter, Joseph A. 1988. *The Collapse of Complex Societies*. Cambridge: Cambridge University Press.
- Walras, Léon. 1874. *Éléments d'Économie Politique Pure*. Lausanne: Corbaz.
- Walras, Léon. 1954. *Elements of Pure Economics*. Translated by William Jaffé. Homewood IL: Irwin.
- Williamson, Jeffrey G. 1974. *Late Nineteenth Century American Development: A General Equilibrium History*. Cambridge: Cambridge University Press.

Suggested Readings

- Johnson, Harry G. 1971. *The Two-Sector Model of General Equilibrium*. Chicago: Aldine-Atherton.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River, N.J.: Macmillan. Chapter 16.

Notes

- 1 Whatever boundaries are used to define an economy under study, and usually they are some sort of political boundary, do define the factor supplies within that region, although other supplies of the same factors undoubtedly exist in other regions defined by other boundaries. If interaction between those economies (regions) is an important part of the problem under consideration, the services of those factor supplies can be exchanged through trade in goods; migration of labor or trade in factors themselves could physically move some of the factors between regions. Archibald (2001, 133–134) notes a lack of regional specification in the influential works of Michael Rostovtzeff and Moses Finley as limitations that hindered the emergence of insights that what she calls a regional approach to many ancient historical issues could have yielded. Indeed, explicit consideration of regions, engaging in interactions with other regions, has been considered an important recent approach to ancient economic history (Reger 1994; Manning 2003; Davies 2001, 44–45). All GE models have such regionalization in them at least implicitly.
- 2 Two important theoretical works on the subject, neither particularly recent, are Bénassy (1982) and Franklin M. Fisher (1983), followed more recently by Fisher (2003). Schultz (1975) is more accessible, but less general, focusing primarily on the value of education to people in modernizing traditional agriculture; it nonetheless offers useful thoughts on behavior regarding long-existing equilibria and sudden changes in circumstances.
- 3 Developed by Walras (1874) in French; available in English translation by William Jaffé (Walras 1954).
- 4 Whether these "markets" are formally organized markets or just loose collections of people who have demands for the same goods and a general awareness of who can be expected to have some willing to sell is immaterial. It is easier to use the term "market" than some considerably longer winded series of qualifications that amounts to the same concept. We have used what is essentially a barter approach to these exchanges of goods, so the presence of money is equally immaterial.
- 5 Alchian and Allen (1969, Chapter 3) offers a more realistic case of exchange without surplus in an example of people in a refugee camp exchanging chocolate bars and cigarettes. See also Radford (1945) for a case of exchange in a World War II prisoner-of-war camp.
- 6 For various treatments, including graphical, of the two-sector model, see Johnson (1971) and (1973, Chapters 7 and 8) and Krauss and Johnson (1975), particularly Chapters 1–3.
- 7 I have used a number of references for this short section: Quirk and Saposnik (1968, Chapter 3); Nikaido (1968, Chapters 1 and 5); Arrow and Hahn (1971, Chapters 5–9); Debreu (1982); Mas-Colell (1985, Chapter 5); Aliprantis *et al.* (1990); Mas-Colell, Whinston, and Green (1995, Chapters 17–19); McKenzie (2002, Chapter 6); (2004, Chapter 5); Moore (2007, Chapter 8).
- 8 If a model contains more relationships than unknowns, it is called overdetermined, and there

can be multiple solutions, some depending on nailing down values of some of the unknowns. A model containing more unknowns than relationships to explain them is called under-determined and has no solution until additional relationships can be specified.

- 9 Prominent among the early applications are Kelley, Williamson, and Cheetham (1972); Kelley and Williamson (1974); Williamson (1974); Kelley and Williamson (1984). Kelly, Williamson and Cheetham, Appendix A, pp. 333–339, contains a proof of existence of equilibrium (as well as uniqueness and stability) for the class of GE model

with labor market equilibrium reached by varying the urban unemployment rate, using a method based on fixed point theorems.

- 10 A “small” country in international trade theory is one that is too small to affect world prices and hence takes its import and export prices as given.
- 11 I have conducted what might be called a poor man’s general equilibrium modeling, but solving only for changes of endogenous variables in response to changes in exogenous variables rather than for levels of those variables: Jones (2000, Appendix B, pp. 189–212).

Public Economics

6.1 Government in the Economy: Scope of Activities, Modern and Ancient

As an agent in the economy, government conducts a number of activities – or equivalently, makes expenditures. To support these activities (that is, finance them), it must raise revenue, in either monetary or “real” terms. This pair of inseparable facts raises two questions: What activities? And, how to raise the revenue?

“Which activities” has been answered differently by different governments throughout history (and probably prehistory) as well as at any one time. Certain types of activities have characterized most governments: defense and judiciary have been pretty widespread. Beyond that, there has been and is considerable variation, with government participation in the production and supply of goods in some societies that private agents provide in other societies. How can a society select which activities to assign to its government? Ideology, of course, plays an important role, but technical characteristics of goods, and occasionally characteristics of production, are important considerations influencing the potential or actual efficiency of provision by either government or private agents.

“How to raise the revenue” has two major options: profits from government production of primarily private goods (we explain that term in the following section) and taxation.¹ The former method has limited scope for raising net revenue. Production costs must be subtracted before “profits” are cleared, and the government producer still has to sell the goods, which leaves revenue subject to consumer preferences, and possibly to competition from other suppliers. These limitations leave taxation as the principal revenue-raising device.

Government is not a monolithic “black box,” but is composed of distinct agencies and agents ranging from royalty to politicians to bureaucrats to employee-“worker bees.” There are many different options for organizing these agencies and agents, differing importantly in the degree of centralization of decision making involved. Regardless of the statutory degree of centralization, effectively there will be considerable fragmentation of decision-making authority, with varying degrees of coordination and competition among agencies. Even with complete centralization, a dictatorial ruler cannot possess all the information required to make all decisions across all agencies; she is thrown back, in the first instance, to the use of agents, for whom she

is the principal. The agents have better access to much agency-specific information than does the ruler as principal, and many of the ruler's organizational decisions embody efforts to protect her own interests – the state's interest – in such an environment of asymmetric information. Each organizational form presents a set of incentives to the full range of government agents. The extent to which the incentives given to government agents lead them to actions that enhance or conflict with the wellbeing of the society is an important characteristic of each set of incentives.

We will use the terms “positive” and “normative” to refer to types of questions posed in this chapter. Positive analysis refers to studying the effects that particular actions or events have. Normative analysis involves evaluation of the effects, deciding whether people are better or worse off after some change or in one of several alternative situations. An example of a positive question is, “What are the effects on labor supply of a tax on income?” A normative question would be, “What is the optimal income tax rate?” To even think about “optimal,” much less decide about it, we need some intellectual machinery to assist in the ethically laden effort of comparing the wellbeing of different persons. Governments routinely make such comparisons and make choices based on them, both usually to a large degree implicitly. The branch of economics devoted to the study of this subject is called welfare economics (it has no direct relationship to public activities commonly called welfare programs, such as assistance to the poor, the sick, and the elderly).

At this point, it is difficult to avoid introducing some technical lexicon. The first piece of such jargon is the concept of Pareto efficiency, which refers to an allocation of resources throughout society such that no rearrangement of resources could be made to make one person better off without making another worse off. The first theorem of welfare economics says that a competitive equilibrium (all marginal rates of transformation in production equal the corresponding marginal rates of substitution in

consumption) is Pareto efficient. In the further reaches of jargon – which itself will actually turn out to be quite useful – a “first-best” situation is a Pareto optimum; stated alternatively, it is Pareto efficient. While the term “first-best” may seem redundant, it is a useful referent to the best situation we can hope to achieve, against which are contrasted “second-best” and “third-best” optima, situations in which the equalization of some of the marginal conditions ($MRT = MRS$) cannot be satisfied. This is an efficiency theorem; it says nothing about the desirability of the efficient outcome, which is contingent upon the initial distribution of resources among individuals (that is, the distribution of income and wealth).

The second theorem of welfare economics states that any desired Pareto-efficient allocation can be achieved as a combination of a competitive equilibrium combined with a set of “lump-sum” transfers. The significance of this theorem is that society can reach any Pareto efficient allocation of consumption by (i) having profit-maximizing, price-taking (recall the conditions of perfect competition) producers carry out production and (ii) redistributing the output in ways that are not influenced by actions that individuals can take. Having said this, we should expand a bit on what “lump-sum” redistribution implies. A lump-sum tax is a tax that the statutory payer of the tax cannot escape by altering his behavior (or any alterable personal characteristics), short of leaving the society. The classic example of the lump-sum tax is the poll tax, an identical amount of tax per “head.” While poll taxes are considered regressive, inasmuch as an identical tax would be a larger percentage of a poor man's possessions (income or wealth) than of a rich man's, it has the ideal efficiency characteristic that nothing that the taxed individual can do can shift the tax onto another. He can change his consumption pattern, he can pretend to be poorer or richer, he can go into another line of business – anything he does will not lessen his tax obligation.

Taken together, these two theorems form the basis for analyzing issues of fairness, interpersonal

and intergenerational equity, and ethically motivated redistribution of consumption in a society. These are exactly the things that governments engage in, directly and indirectly, through their expenditure and revenue raising activities. The familiar tradeoff between efficiency and equity – the size of the pie versus how it is sliced – is contained in the same analytical apparatus. In the spirit of revealed preference, major parts of the ethical systems of societies, at least as they are represented in their governments, can be inferred from the actions of government. This concern with distributional ethos will not be the central focus of this chapter, but it does form a relatively constant backdrop to the positive analyses presented below and is one of the twin pillars of the normative analyses.

As the first analytical topic, we turn to the concept of the public good, which is defined relative to private goods. The analysis of the first five chapters has dealt with private goods, although we didn't make a big fuss over it. The difference between the two types of good will be quite clear. Defining what "publicness" in a good is motivates studying how government raises revenues to finance public goods. Accordingly, the third section is primarily devoted to the analysis of taxation; the last subsection of that section discusses other methods of raising revenue. Section 6.4 introduces the general issue of what policy actions are appropriate toward "well-functioning" areas of the economy when some other areas of the economy "don't work so well" – another topic in the interconnectedness of different parts of an economic system. Section 6.5 treats government provision of both private and public goods, including investment in the construction of durable public goods and methods for deciding how much is enough. Some government activities are more admonitory and restraining than actively creative; section 6.6 addresses regulatory activities of government, in contrast to the budgetary issues treated till this point. The final section turns to models of the behavior of government as a set of agents itself.

6.2 Private Goods, Public Goods, and Externalities

6.2.1 Private goods

Private goods have the characteristic that one person's consumption of them generally pre-empts another's consumption. If I eat an apple, that's one apple that you can't eat. There may be other apples, but the one in question is taken. This characteristic is called *exhaustability*. You will occasionally see the characteristic referred to as "rivalry in consumption," or "rivalrousness" in consumption. They all mean the same thing. Another distinct characteristic of private goods is that it is easy to keep people who don't buy them from consuming them – short of stealing them, that is. This characteristic is called *excludability*. With these two characteristics, prices are effective allocation mechanisms, both for dispensing of goods once they are produced and for telling producers how much to produce. Recall that we can aggregate individual demand curves for a good (that is, the demand curves of different individuals for a particular product) by adding them horizontally: at price p_0 in Figure 6.1, Smith will take two units, Wilson will take four and Dumbrowski three, making in all nine units. The aggregate demand curve coincides with Dumbrowski's demand curve from its vertical intercept

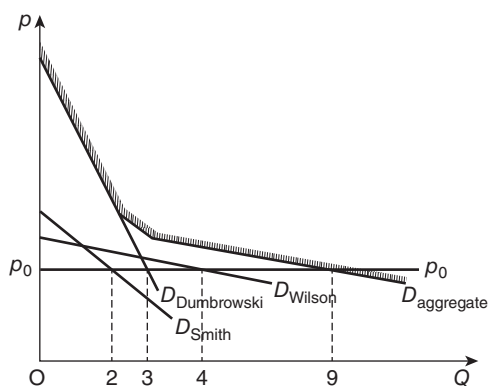


Figure 6.1 Market demand for a private good.

down to the vertical intercept of Wilson's demand curve, at which point it becomes the horizontal sum of the different individual curves.

6.2.2 *Public goods*

Public goods are best defined in contrast to these two characteristics of private goods. First, it is possible for many people to enjoy the consumption of a public good without diminishing the quantity of the public good available for others to consume. National defense is one of the best examples of this characteristic: one person's consumption of national defense does not subtract from another's. A judiciary system, with its laws protecting citizens and identifying their rights, is another example of inexhaustibility. Both of these examples are what could be called "pure" public goods, in that their inexhaustibility genuinely seems to have no limits. A somewhat less pure public good would be a bridge over a river or a court in a judiciary system. The smallest size bridge that will stand architecturally is far bigger than any one individual would want to consume; even a single-lane footpath most individuals would not want to pay for all the time – just when they use it. The large minimum size relative to individual demands makes it score high on "publicness," but, with enough traffic on it at one time, it can get congested. An individual court, or even all the courts within a judiciary system could get clogged up with cases if the proportion of cases to court personnel gets too high. Let's keep the bridge and court examples in mind while we turn to the other major characteristic of public goods, nonexcludability.

The other contrast to private goods offered by public goods is that it is difficult to exclude nonpaying individuals from consuming them. National defense and the judiciary are excellent examples. Even if one individual paid nothing at all – say, through successful evasion of all taxes – he could not be excluded from enjoying the national defense and the legal system provided by the government. Whether the consumption of a particular good can be kept from people who do not pay for it really is a matter of how much it costs to monitor consumption and

enforce payments. In the case of the bridge, it might be thought that some system of fees would be useful for both rationing use and defraying the cost of provision. Recall that setting price equal to marginal cost is desirable for efficiency, but recall also what we said about the minimum size of the bridge relative to any one consumer's demand for it. Such large initial costs imply a declining marginal cost curve, which will be below the average cost curve as long as it is declining, and so marginal-cost pricing will not cover total cost. Any price that covers total cost will be above marginal cost, but marginal cost is what each user contributes to the total user cost of the good, so we would have to charge each user more than the marginal cost of his use to cover total cost. A private supplier, who had to cover costs directly from sales or user fees, would have to charge a price in excess of marginal cost, which would lead to under-utilization (users would set marginal benefit equal to the price they had to pay, so $MC > MB$ would lead to underutilization). While we could obtain our supply of bridge services from private providers, there is something to be gained from public (government) provision since government has more means to secure revenue than do private suppliers.

We can see the consequences of the nonexcludability for the aggregation of demands for a public good. Let a large number of consumers each maximize their individual utilities, which are a function of their consumption of a single private good plus a nonexcludable public good: $U^i(X^i - t_i p_G G, G)$, where X_i is the endowment of the private good, the price of which is normalized to one; p_G is the price of the entire public good G , and t_i is a share of the public good price, such that $\sum_i t_i = 1$. Recall the definition of a first-order condition: we (the agent in the optimization problem) vary the "instrument variables" in the objective function (the utility function in this case) to find the value of each instrument (X_i and G in this case) that gives the largest value of the objective. Varying X_i and G in this manner and equating first-order conditions yields the result that the marginal rate of substitution for each individual i is $(\Delta U_i / \Delta G) / (\Delta U_i / \Delta X_i) = t_i p_G$.

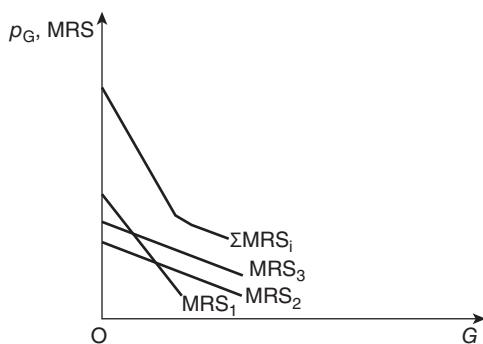


Figure 6.2 Market demand for a public good.

Now, sum this relationship over all individuals: $\sum_i \text{MRS}_i = p_G \sum_i t_i = p_G$ since the t_i sum to 1. Because each individual consumed the full amount G of the public good, the optimal quantity to supply of it is the sum of the amount each consumer is willing to contribute. For each consumer that amount is his marginal rate of substitution of the public good for the private good, which itself amounts to a quantity of the private good each consumer is willing to surrender for his consumption of the public good. In Figure 6.2, begin with the MRS curves for three individuals. Each point on an MRS curve, measuring vertically from the horizontal axis, is the amount of the private good the consumer is willing to offer for an additional unit of the public good. Recall that this is what a demand curve shows; these are demand curves for a public good. We sum these vertical measures to obtain the aggregate demand curve for the public good, $\sum \text{MRS}_i$.

Note that the public good concept does not imply that consumers receive equal benefits from those goods. There is a subtle difference here: while each consumer consumes the same quantity of a pure public good, consumers can receive different marginal utilities from that consumption. All individuals consume the same amount G of the public good, but that does not imply that $U^i(G) = U(G)$ for all individuals i ; nor does it imply that $(\Delta U^i / \Delta G) / (\Delta U^i / \Delta X_i)$ is the same for all individuals.

The sleight of hand in this demonstration of the demand for a public good involves the

variable t_i , about which we were rather reticent. The t_i represents shares of the total cost (price) of the public good that each individual is really willing to pay for. They are unobservable, and once each individual understands the payment system – that is, that he is going to pay the share t_i of the price p_G because that is his true valuation of it – each recognizes his own incentive to understate his demand for the good. If everyone else were truthful, and one consumer understated his preference for the good, he could enjoy just an infinitesimally smaller amount than he could have had if he had spoken truthfully, but at substantially less cost to himself. This is the free rider problem, which can become quite serious when everybody understates his own demand. These individualized tax shares, $t_i G$, are called Lindahl prices, and they satisfy the necessary condition for a Pareto-efficient supply of the public good. The only problem with them is that learning what the t_i are is literally an impossible task: there is no mechanism that will reveal these preferences truthfully.² We will return to the subject of mechanisms for making collective, or social, choices such as these in section 6.5.

Before leaving public goods, we should emphasize the distinction between government (or public) production of private goods, which has not been the subject of this discussion, and public provision of collective goods. We recognize the potential importance of public production of private goods in the study of ancient eastern Mediterranean societies, and we address that topic in section 6.5. Further, the borderline between private and public provision of a number of public goods may shift over time according to technologies available and for other reasons. For example, it appears that institutional (public) controls governed the construction and use of major streets in Old Babylonian cities, while some mix of private actions and public adjudication appears to have combined to govern the construction, use, and maintenance of smaller streets and alleys (Keith 2003, 62–63).

6.2.3 Externalities

The nonexcludability characteristic of the pure public good carries over to some private goods

and actions. Production of metals involves burning fuels and oxidizing various elements in ores, which can produce smoke and noxious fumes, bad to smell, bad for the health, and possibly damaging to other production and consumption (or household production) activities in the area. The classic example of interacting productive processes is a smoke-producing metals factory dropping ash on a nearby laundry, requiring rewashing or other protective measures by the latter. In this case a negative externality exists. But consider the case of health care, in which treatment of one person for a contagious disease reduces the likelihood that others will contract the same disease. If the person with the ailment considers only his own marginal benefits from treatment, he will seek treatment to the point at which his marginal cost (the payment for the treatment) equals his marginal benefit, ignoring the benefits to others. In many cases, there is no effective method for “passing the hat” – it may be difficult to both identify the beneficiaries and persuade them to reveal their true valuations of the reduced likelihood of their contracting the same ailment.³ At any rate, this is an example of a positive externality.

Externalities may derive from both production and consumption activities. An externality may be said to exist when either a consumer's utility function contains some consumption not under her direct control or producer's production function contains some factor of production whose supply she does not control. Thus, in the utility case, suppose we have two individuals we call A and B. What individual B does with his own consumption of good j enters A's utility function with the level of good j that individual B chose: $u^A = u^A(x_1^A, x_2^A, \dots, x_n^A, x_j^B)$. Individual A has no say over the amount of good j that she consumes: she gets individual B's choice. If she likes good j , she might like to have more of it than B is providing for himself, but she has no direct, market means of getting B to consume more of it – she might not even know who individual B is. If she doesn't like it, she also has no means of getting B to reduce his consumption of it. What does it mean to say that A “likes” or “dislikes” her incidental consumption of good

j ? If A's marginal utility of good j , at the level of j chosen by individual B, is positive, A feels better off with the externality than without it; she might even be willing to pay something to have more of it. If A's marginal utility of good j is negative, she would be able to improve her wellbeing by paying some amount to get the level of good j reduced – an amount equal to the negative of her marginal utility of good j . If B's consumption of good j is in A's utility function but it has a zero marginal utility, either because B isn't consuming enough of it to be noticeable to A or because he's not consuming any of it at the moment, the externality still exists, or because he's consuming the amount that turns A's marginal utility from positive to negative, hitting zero on the way, but it can be called an inframarginal externality. In the other two cases, with either positive or negative marginal utility flowing from the level of j chosen by B, there is a marginal externality.

Suppose A's marginal utility of B's choice of good j consumption is negative. If we require B to either reduce his consumption of j or cease and desist altogether, A is better off, but B's wellbeing is reduced without compensation. The externality still exists, and we have just transferred the negative consequences of it from A to B, and there is no reason to think that this is just. Externalities are symmetric.

To help us dig a bit deeper into the cause of externalities, let's give a concrete example to the case of A and B's consumption of good j . Suppose B smokes a pipe, and A likes a certain amount of pipe smoke, but beyond that amount becomes saturated then begins to feel positively sick. If B continues to smoke into this sickening range, A feels abused, but if B is forced to cut back before his marginal benefit equals his private marginal cost, he feels abused, each with some justification.⁴ The problem is that smoking the pipe uses the air as a disposal system, but neither party has property rights to the air, so it remains unpriced. B's smoking makes a *de facto* exercise of a property right, but in fact he has paid no one for it and is not paying the marginal cost of any damages he inflicts, in this case, A's disutility of pipe

smoke. To simply assign the property right to air to A and let her charge B for smoking may be felt to solve the externality problem in some sense, because B must now pay a cost for the negative externality he inflicts by using A's air, in addition to the marginal cost of a load of tobacco (its purchase price, ignoring the cost of the pipe, which is a sunk cost; abstracting from health effects). However, this solution confers a windfall on individual A for no good reason. Following the symmetry of the problem, we could have assigned the property right to the air to B and let him charge A for not smoking.

Figure 6.3 illustrates this problem, measuring A's marginal benefits in the upper half of the diagram and measuring B's negatively in the lower half. The straight line below the pipe-smoking axis (above it from B's perspective) is B's marginal cost of pipe smoking, which could be simply the market price of a load of tobacco. At the point B^* on the pipe-smoking axis, B's marginal cost of pipe smoking equals his marginal benefit, and in the absence of any forced reductions or trades, he would choose that consumption level. A, however, reaches her saturation point with B's pipe smoke at point $A^* < B^*$. At the level of pipe smoking represented by A^* , B's marginal benefit from smoking is still well above

his marginal cost. At B^* , A's marginal utility of B's pipe smoke is far into the negative range (disutility). We draw the curve MB_B^{net} by subtracting out the constant marginal cost of the pipe tobacco and find that at the level of pipe smoking identified as T, B's net marginal utility is equal to A's marginal disutility. A would feel herself better off if she paid B an amount equal to his net marginal utility at point T, and B would be satisfied with the trade. At this point the marginal damage caused by B's pipe smoke is equal to his marginal benefit. In other words, consumption level T is the socially optimal level of B's pipe smoking, and A and B have arrived at a Pareto-optimal solution to their externality problem. We can say that the externality has been internalized. The externality still exists – A still has an inframarginal disutility from B's pipe smoke, but neither party can do better without harming the other. To the right of point T in Figure 6.3, there was room for Pareto improvement since one party could be made better off without the other being made worse off. To the left of T, A could be made better off but only at the cost of B's being worse off.

The general principle about internalizing externalities that emerges from this example is that it is inappropriate (counterproductive, inefficient, not welfare-enhancing) for the generator of the externality to compensate the "victim."⁵ This may seem counterintuitive, but it is a very robust result in the theoretical analysis of externalities. "Why not?!" is a logical question. First, compensating the "victim" can give individuals incentive to expose themselves to the externality in order to collect the compensation. Second, and related to the first reason, compensation would reduce the incentive of the consumer of the externality (the fellow we were calling the "victim") to adjust his behavior so as to avoid it or at least reduce his consumption of it. In the case of the metal factory and the laundry to which we referred in passing above, it could prove much less costly to the economy as a whole for the laundry to relocate to avoid the ash fall than it would to move, say, an iron foundry or

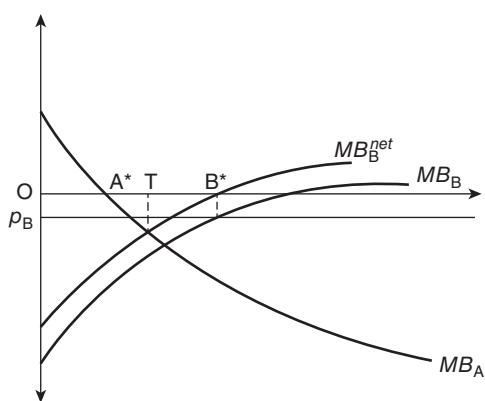


Figure 6.3 Marginal benefit of reducing an externality.

a steel mill. The designation of smoking and nonsmoking sections of public facilities such as restaurants and airports in recent years, prior to the more recent outright bans, particularly in the United States, is an example of separating the generators and recipients of an externality. A particularly smoke-sensitive individual might point out that he or she usually can still smell the smoke under this separation policy: why don't the public authorities just put the smokers in a sealed room or somewhere even further away from the nonsmokers? So what if they have to walk a mile or so and have a lousy view when they get there? We've already answered this query in Figure 6.3 and in the example of B's pipe smoking: at the Pareto optimum (or even less satisfactory allocations; Pareto optimum is just the best that can be achieved), there is still some amount of the externality, but obtaining less of it would reduce overall wellbeing, including that of the smokers.

There is considerable conceptual similarity between consumption externalities and production externalities, particularly in the conditions for internalization. Nevertheless, there is sufficient scope for additional insight to make it worthwhile dealing directly with production externalities. Production externalities can operate directly between producers. One type of production externality is known as the unpriced factor situation; the output of one production process affects the input set to another production activity. The classic example is honey and apples: the honey bees get pollen – their food – from the apple trees, and there is no way the apple grower can stop the bees or charge the beekeeper for their food. (A symmetric effect may benefit the apple grower as well if the bees provide pollination services for the apple trees, raising the productivity of the apple orchard.) The beekeeper doesn't have to pay the apple farmer for the services of his trees, and in fact has no way to keep his bees from visiting the apple grower even if he wanted to, short of moving his bees out of flying range of the trees. However, in all but a fortuitous case, the marginal productivity of the apples in the beekeeper's production process will exceed their marginal cost to him, and he

should be willing to pay to get more apple services since he would profit from the additional inputs. The beekeeper would appear to be producing under increasing returns to scale, although when the services of the apple trees are accounted, constant returns to scale in both activities will still occur. The production functions look like the following: Apples = $f(\text{Land}_a, \text{Labor}_a, \text{Bees})$, and Honey = $g(\text{Land}_h, \text{Labor}_h, \text{Apples})$, with the apples unpriced, inexhaustible (the bees don't eat the apples), and nonexcludable.

Another type of production externality is the atmosphere externality, in which the output of one activity shifts the entire production function of another activity. James Meade's (1952) example was timber and wheat, with the afforestation of the timber industry increasing rainfall available for the wheat. The production functions are: Timber = $f(\text{Land}_t, \text{Labor}_t)$ and Wheat = $g(\text{Land}_w, \text{Labor}_w) \cdot h(\text{Timber})$. In this case, the output of the timber industry increases the productivity of both factors in the wheat industry, having the effect of giving increasing returns to scale in wheat production. The atmosphere created by the timber industry is freely available to all producers in the wheat industry; they have to do nothing to receive it. The payments to the factors employed in timber are below their marginal social values, however, and it would raise total output across both industries to shift some land and labor from wheat to timber production, although it would take a tax or subsidy to accomplish that, and in practice that could prove quite difficult.

These two types of production externality have dealt with the effects of the outputs of one industry (or activity) on another industry (activity). It is possible that the employment of some factor in one industry could confer either positive or negative externalities on the output of another industry, or even on the productivity of some factor in another industry. The structure of possible production externalities is quite flexible, but in all the cases, the externality will cause a divergence between private and social marginal cost and productivity of factors, calling for a reallocation of factors among activities. With decentralized ownership and operation

of the interacting activities, such a reallocation would require confronting the factor suppliers with a different set of factor prices, most likely through a combination of subsidies and possibly taxes, the details of which we will not delve into here.

Let's turn to negative production externalities and methods of reducing them. The two principal methods for moving externalities to their optimal levels (or at least closer in that direction than they would be in an uncontrolled situation) are taxes and direct quantitative controls (emissions limits, required equipment, and so on). Figure 6.4 shows the application of a tax to an externality. The demand and supply curves for the good whose production generates an externality are shown as D and S . With no intervention to control the externality, the equilibrium price and quantity are p and q . The curve above the supply curve, labeled MSC , is the marginal social cost curve, which includes the value of the damages or disutility imposed by the production of the good. The optimal quantity of this good, q^* , is determined by the intersection of the marginal social cost curve and the demand curve, point a in the diagram. The cost / price difference between the marginal social cost and the private supply price at quantity q^* is the distance ab . A tax equal to ab on each unit of this good will move the supply-demand equilibrium from (p, q) to $(p^* + t, q^*)$, leaving p^* as the price received by the producer.

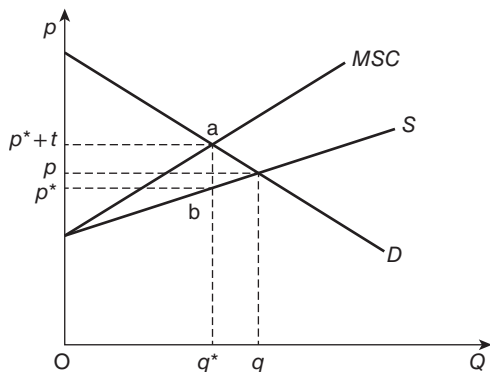


Figure 6.4 Market view of a negative externality.

Figures 6.5 and 6.6 offer two alternative perspectives on the control of externalities. Figure 6.5 is a generalization and simplification of Figure 6.3, showing the relationship between the level of the activity generating an externality and its benefits and costs. The marginal damages curve is rising, much like a private cost curve would. This slope indicates that the disutility caused by additional units of the activity that causes the externality gets worse as more of the activity is conducted. The marginal damage curve could be flat, indicating constant costs (not no costs). It is not entirely inconceivable that this curve could have a negative slope, in which case it would have to cut the marginal benefits curve from below to justify attempts to contain the activity.⁶ The downward-sloping marginal benefit curve derives from utility theory. A social optimum

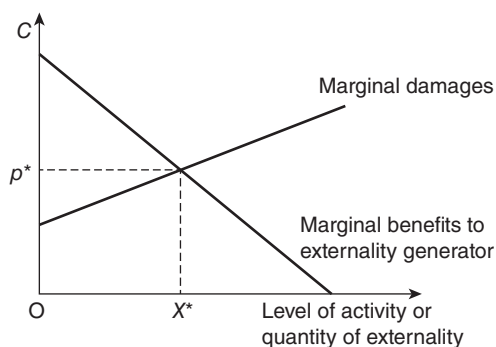


Figure 6.5 Equilibrium level of an externality.

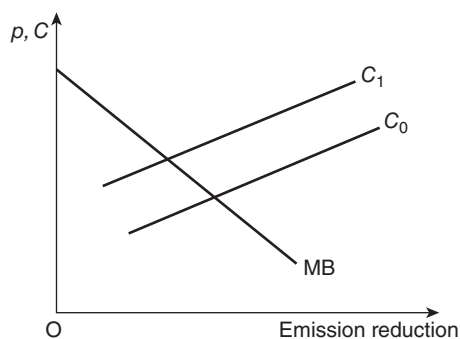


Figure 6.6 Control technologies and the choice of an externality level.

would be achieved by equating marginal damages to marginal benefits. Since there is not a market to guide this equalization, the agents involved in the situation are left with either individual bargaining, as in our example of the pipe smoking, or via publicly imposed taxation or quantity restrictions. Figure 6.6 turns around the variable measured on the horizontal axis of Figure 6.5, measuring the *reduction* of the externality emission; points farther to the right in this figure represent greater efforts to control the externality. The downward-sloping marginal benefit curve in Figure 6.6 refers to the marginal benefits of reducing emissions, not to the marginal benefits of consuming the good that released the emissions to someone else. Its downward slope indicates that the first units of emission reduction are the most valuable; the more we control the emission of the noxious substance, the less it is worth in the opinions of the people represented in the diagram to reduce it by more. Turn back to the example of the smokers in the restaurant; how much would it be worth to you nonsmokers to get rid of the smokers altogether? Evidently not as much as it would cost. The emission-control cost curves, C_0 and C_1 , slope upward just like an ordinary cost or supply curve would: the more we control of a substance that, if turned loose, would produce an externality, the more it costs. Curve C_0 represents a lower-cost control technology than does curve C_1 . Again, equating the marginal benefit of emission reduction to the marginal cost of same will deliver the efficient quantity of control.

The reader would be correct to infer that establishing the MSC curve would be more demanding of information than would the supply curve. For one thing, nobody really needs to know the industry supply curve to conduct business; individual producers make a few guesses, assumptions, and calculations, and the forces of incentives finding their way through scarcity will cause outcomes to more or less follow patterns that look like a supply curve, but the real value of a supply curve is for people who want to study the economy rather than for those who want to participate in it. But aside from the

philosophical note these sentiments express, the fact of the matter is that for public agents to participate in the economy inasmuch as they impose a tax on this externality, they really do need to know the value of the damages, at least in the “region” (the price-quantity region) of the *laissez-faire* equilibrium. In contemporary environmental policy, this information must be contributed by the relevant scientific industries, and is subject to considerable public debate, most of it not disinterested in the least. The precision of the narrow lines in these diagrams can lend an impression of greater precision in the application of contemporary environmental policy than is the case – and could lead the reader to think that spending scarce time thinking about externality policy is a waste of that scarce and valuable time for people who spend their lives thinking about ancient economies and societies. However, think about some of the examples of externality-generating activities in the ancient world. Ore smelting, metal casting, and pottery firing are three obvious cases in which the ancients could hardly have failed to notice impacts on other parties, including even probably health effects on workers involved in metals production after an extended time of inhaling fumes. The locations of these activities in and around the ancient cities may give some hints of ancient public policy toward them. The archaeological remains of apparent metal smelting in a Late Bronze Age temple at Enkomi on Cyprus (Dikaïos 1969, 76–81; Courtois 1982, 175; Stech 1982, 105–106, 108–109; Tylcote 1982, 89–92, 102, Fig. 5) may have represented an unusual locational choice, despite contemporary inferences about the apparent sacredness of metal working at the time. The remains of kilns, particularly in heavily urbanized parts of the ancient eastern Mediterranean region, appear to have been located toward the outskirts of built-up areas. Holleran (2012, 59) notes the location of potteries in Rome toward the outskirts of the city, around the present-day Vatican, in consideration of the risk of fire associated with ceramic processes. Tanneries in Rome were also located toward the outskirts of the city (Holleran, 2012, 58) for the obvious reason of their smells. In less

densely populated areas in Greece, there may have been less of a tendency to avoid proximity to settlements because fewer people were nearby to be affected. In agriculture, one farmer's actions frequently could have affected his neighbor's farming in undesirable ways: the cattle get loose and trample a neighbor's crop, a ditch dug on one's own property diverts water in an undesirable direction from his neighbor's perspective, failure to care for crop land on a slope dumps soil and rocks on a neighbor's property. Plenty of these cases probably showed up in court, and records of some of them may remain in the tablets or papyri.

Before leaving the subject of externality, we feel obliged to mention the concept of pecuniary externality, which stands in distinction to the type of externality we have been discussing so far, which depends on the functional structure of production and utility (that is, the variables that are in production and utility functions) and are called technical externalities. You will occasionally see reference to pecuniary externalities as situations in which one agent's action raises the cost of another agent's activities, but does not directly affect that agent's consumption or production levels. Frequently the actions of markets can cause such price changes – indeed that is what markets are supposed to do. The entire concept of the externality (you can call it the technical, or true, externality) involves the absence of markets to price the transactions of certain goods. The pecuniary externality is a bogus concept and can serve mostly to confuse the unwary. If you stumble across it in your reading, go on to something else.

6.3 Raising Revenue

The first four parts of this section are devoted to the analysis of taxation. To a large extent this reflects the relative importance of taxation in raising public revenue. Government debt policy will be not treated here, but even sovereign borrowings have to be repaid (at least enough to keep debt, and the interest on it, from becoming infinite), and that will come mostly out of future

tax revenues simply because of the relative dominance of taxation in public revenue sources. The first subsection provides a background on tax philosophy; types of taxes, showing the range of tax instruments available, many of which probably were used by ancient governments, and their informational and administrative requirements; and an introduction to the analytical vehicle for thinking about equity versus efficiency, the social welfare function. The second subsection introduces the analysis of the effects of taxes on particular elements of activity. The third subsection treats the issue of who actually pays various taxes, a subject called tax incidence. Both this and the previous subsection are topics in the positive analysis of taxation. The fourth section introduces an important normative topic in taxation: optimal taxation. Skirting the precise content of optimal taxation just a bit, since what's optimal depends on preferences as well as technologies, different societies may find different tax systems best for them. The last subsection deals very briefly with nontax means of raising public revenue.

6.3.1 *Taxation 1: rationales and instruments*

Equity versus efficiency

There are two major philosophies of taxation (or more generally, the combination of taxation plus receipt of benefits from government expenditure): the benefits and ability-to-pay principles. While they are contrasting philosophies, it is common to find both rationales guiding the opinions of both individuals and societies as they attempt to resolve the tensions inherent in any tax policy. The benefits principle looks to efficiency in the supply of public goods. In its starkest form, it would have those who benefit from government services pay for them. This seems eminently reasonable until we consider the fact that many individuals in any society are quite poor, at least relative to other members of the society. Tempering the potential for what would strike many as injustice inherent in the benefits approach to taxation is the ability-to-pay principle, which recognizes both inequalities in

wealth and income and the mandatory character of taxation. It is one thing to let the poor choose not to purchase / consume items they cannot afford, but quite another to force them to pay taxes that could leave them with not enough to eat.

Two widely considered concepts of fairness in taxation are called vertical and horizontal equity. Horizontal equity involves the equal treatment of agents who are identical in economically relevant characteristics and endeavors to avoid differential taxation on the basis of “irrelevant” characteristics. For instance, many contemporary societies would feel uncomfortable taxing people of different religions differently, although that has been a subject for varying opinions over the centuries and across the continents. Another example that may transcend cultural differences more completely would be a dictum that people with identical circumstances should be taxed identically, but there are difficulties in deciding on whether *ex ante* or *ex post* circumstances are the relevant targets of an equal-treatment-of-equals policy. For example, equal treatment of men and women in social security taxation ignores the fact that women tend to live longer than men; should “equal treatment” mean they are taxed the same although the women can be expected to receive benefits longer or taxed differently on the grounds that, actuarially, the women can expect to receive more in social-security benefits? Vertical equity focuses on the proper treatment of individuals who differ on characteristics that are thought relevant to their taxable capacity. This principle, naturally, comes from the ability-to-pay philosophy. Both principles involve interpersonal judgments about wellbeing, which all economists accept as outside the scope of economic science, although economic methodology may help clarify the questions and choices.

Over the past hundred or so years, there have been quite a few recommended implementations of both principles (for example, absolute benefit, marginal benefit, relative marginal benefit, and so forth), but the relationships of the various measures to one another have taken considerable time and energy to clarify. During this century, the concept of the tradeoff between the efficiency

of the benefits principle and the equity of the ability-to-pay principle has emerged as a potentially unifying analytical device. Identifying such tradeoffs has fallen clearly within the scope of science, but recommending one over another on any scientific basis has become recognized as a matter of ethics outside the scope of even normative economics because it involves interpersonal welfare comparisons.

The social welfare function

The social welfare function is the analytical tool used to combine the study of efficiency and equity. The simplest form of social welfare function (swf) is $W = W(U_1, U_2, \dots, U_n)$, where the U_i are the utilities of the n individuals in the society. In this sense, the swf is individualistic in that it reflects the wellbeing of each individual. Different functional forms of the swf reflect different philosophical and ethical views. If we specified $W = U_1 + U_2 + \dots + U_n$, we have what is known as the Benthamite social welfare function, which is simply the “greatest sum of the happinesses.” John Rawls’s proposal for the just society, which was that the distributional goal of the just society should be to maximize the welfare of the poorest member of society, would be $W = \min_i(U_i)$. A quite general form that contains both the Benthamite and Rawlsian functions as special cases is $W = (1/(1-\nu)) \sum_i [(U_i)^{1-\nu} - 1]$, in which $\nu = 0$ generates the Benthamite case and $\nu = \text{infinity}$ gives the Rawlsian (Atkinson and Stiglitz 1980, 339–340).

A much more specific functional form can be constructed to reflect the sacrifice in efficiency a society is willing to tolerate to achieve a particular amount of income redistribution.⁷ Let δ be the proportion of an income transfer the society is willing to have “evaporate” during the transfer from a wealthier individual to a poorer one. Then assign welfare weights β_i (think of weighting the utilities of different individuals in the Benthamite form of the swf) having the relationship $\beta_i/\beta_j = (1-\delta)^{Y_j - Y_i}$, where individual i ’s income is greater than individual j ’s. Now, suppose that the weights β_i depend only on income, Y_i ; we can write a functional form for the welfare weights $\beta_i = aY_i^e$. Substituting this form into the

weight ratio, we get $(Y_j/Y_i)^e = 1 - \delta$. Suppose individual i has twice the income of individual j ($Y_i = 2Y_j$). If society is willing to lose half of the income it tries to transfer, $\varepsilon = 1$; if $\delta = 3/4$, $\varepsilon = 2$. Alternatively, suppose that the society under consideration is concerned about dimensions of poverty other than relative incomes; specifically, the absolute level of income relative to some specified minimum is of particular concern. Then the functional form for the weights β_i could take a form like $\beta_i = a(Y - \bar{Y})^{-\varepsilon}$, in which $\bar{Y}$ is the minimum subsistence level of income. As some individual's income Y_i approached the subsistence level $\bar{Y}$, the value of the weight β_i would approach infinity. One problem, of course, with implementing a redistribution system based on the concept of a minimum subsistence income is that in any society, many families and individuals actually manage to live on less than the specified minimum.⁸

Clearly, most governments, and certainly entire societies, do not think in terms of a social welfare function. The justification for using the concept, once again, is that it can capture succinctly some of the important ethical judgments that a collectivity makes. The social welfare function is used widely in the analysis of tax and public expenditure systems: as analysts, we posit that an agent of the society ("the government") maximizes a social welfare function subject to a number of technological, behavioral, and budgetary constraints, adjusting tax and expenditure patterns to conduct the maximization (we'll see this at somewhat closer range in subsequent subsections). The pattern of taxes and expenditures that maximizes the objective function depends on the form of the objective function – that is, on just what the form of the swf is. Given the maximizing values of taxes and expenditures and knowledge of the constraints, it is theoretically possible to "back out" some information on the form of the swf that could have yielded the "optimal" values of the instrument variables (the taxes and expenditures). Again, as analysts, we can use the swf construct to reflect the preferences of interest groups: specify the swf form that a particular set of interest groups could agree upon, conduct the maximization, specify a slightly

different swf form, maximize again, compare the two sets of optimal instrument values. In yet another interpretation, the swf could be viewed as a positive description of how a particular government operates, recognizing that the swf does not suppose that the government is either benevolent or monolithic.

Taxes: categories, concepts, and specific instruments

It is common to find taxes divided into the categories "direct" and "indirect," and their meanings are not necessarily intuitively obvious. Direct taxes refer to taxes on factors of production – that is, on sources of income. The term "direct" is used to name this category of taxes because the tax rate depends on characteristics of the individual, household, or whatever unit is being taxed. The tax rates can be personalized. Personal and corporate income taxes are direct taxes. Indirect taxes are taxes on commodities (goods and services) – basically, everything else besides direct taxes, because some taxes like the sales tax, while nominally on goods and services, are really taxes on the value of transactions. Indirect taxes are taxes on uses of income, as contrasted to sources. The tax rates generally are invariant to the characteristics of the agent purchasing the taxed items.

An *ad valorem* tax is one that claims a given percentage of the value of the base: for example, a 5% sales tax or a 10% tax (it is more common to hear *ad valorem* used in reference to commodity taxes than to direct taxes, but the principle is the same). Specific taxes are levied in a fixed amount per unit of a base, again usually a commodity. A contemporary example would be, say, a 20¢ per gallon retail tax on gasoline. The tax literature makes common reference to what is called the "lump-sum" tax, the only effective example of which is the poll tax, which is much less common than it has been in previous times. Few cases are actually found of lump-sum taxes, but they make a handy theoretical referent in analytical discussions of income redistribution.

Some taxes, or combinations of taxes, are equivalent to other taxes. For example, a uniform tax on all commodities (including services) is

equivalent to a proportional tax on income. A uniform tax on all commodities is more difficult to achieve than might be thought. Housing is a major expenditure item in the budgets of most consumers; the proper target of a commodity tax on housing would be a tax on the flow of services from a house during a given time period (say, a year). Identifying these flows would be a major effort, adjusting for the hedonic characteristics of different houses, accounting for depreciation and maintenance, deciding how to treat the number of users and specific uses. Each year these characteristics could change, and the value of the flows could change in response to any number of determinants of the demand for, and supply of, housing. It would take an unusual public authority to want to raise part of its revenue in this manner. Once we drop housing from the array of commodities to be taxed, the equivalence of the uniform tax on all commodities and the proportional income tax evaporates.

The example of implementing the tax on the flow of housing services does, however, introduce a number of considerations that will influence the choice of tax instruments by a public authority. At the most general, administrative cost is a primary consideration. The components of administrative cost include the cost of assembling information about the tax base (individuals, commodities, sales, and so forth). Individuals with tax obligations of one sort or another can take actions to alter or disguise their characteristics that affect those obligations: the wealthy can attempt to appear poorer (relevant to an income tax), the successful can attempt to look less successful (relevant to both income and profits taxes), and so on. These efforts can fall either within or outside the tax laws, and the clarity of where actions lie on that continuum will affect the tax authority's costs of collecting the tax. The ready observability of the tax base or tax obligation is an important consideration. If it costs the tax authority, say, 50% of the tax revenue it collects to observe what the tax obligations are, that base is effectively unobservable. Despite the record-keeping ability in the ancient Mediterranean societies, observability of tax bases must

have been an important factor in their choices of both tax bases and tax instruments. One rule of thumb in tax administration is that high rates of taxation make tax evasion more profitable; conversely, low rates make concerted efforts at evasion not worth the trouble. The intelligent tax authority would not like to give taxpayers additional incentives to evade, although cases occur in which a combination of recognized evasion by some groups and popular appeals to tax those groups at even higher rates to compensate for their evasion, raise the evasion rate more rapidly than the tax rate, and the entire tax collapses from near-complete evasion.

Another important consideration to the government considering the mix of taxes it wants to use is the stability of the revenue stream that any particular tax would yield. Is a particular tax base particularly sensitive to regular or irregular cycles? Regular cycles may be easier to surmount than the irregular ones if they occur at sufficiently short intervals. For instance, cycles within a single year, when the year is the temporal basis of tax payments, would pose little problem. On the other hand, a tax on, say, landings of a fish that enters local waters only once every two or three years, even if in considerable numbers, might lend an unacceptable degree of instability to justify imposing a tax on that specific item. Related to the stability of the tax revenue is what is called its elasticity: the extent to which the growth of the tax base parallels the growth of the economy in general. Clearly this consideration is more important in an economy in which growth of income per capita is anticipated over a relatively short period, say 5 to 10 years. For example, while taxes on commodities with low own-price elasticities of demand can be expected to be quite effective at raising revenue (for reasons to be explained in detail shortly), these same commodities tend to have low income elasticities of demand as well (think back to the relations between price and income elasticities of demand in Chapter 3). Consequently, if the income of the community is expected to grow, the tax authority cannot expect to share much in that growth with such a tax.

Finally, but possibly much closer to the top of the considerations list of the tax authority is political acceptability. It is common belief that less visible taxes tend to be less unpopular. While sales and excise taxes are visible, the fact that their costs are typically bundled with the gross prices that consumers pay has raised their acceptability, if not necessarily their popularity. Unpopularity with highly concentrated interest groups can greatly reduce the political viability of particular taxes, or particular provisions of personal and factor-specific income taxes.

Let's turn to some individual taxes. Considerable variation is available in income taxes. We could have, simultaneously, taxes on personal income and taxes on business income (most countries do contemporarily, with the personal and corporation income taxes). The personal income tax could tax different sources of income at different rates, and the tax rate on any particular source of income, or on all, need not be linear: that is, at a fixed rate. The progressive income tax schedule is a well-known example, in which the tax rate rises at higher income levels. Income taxes can offer allowances for certain types of expenditure and fluctuations in income. The latter includes the important allowance of offsets against tax obligations for losses. Taxation of income from capital in a personal income tax is a tax on savings. Losses, of course, will occur primarily in income from capital. In the taxation of income from capital, governments frequently tax capital income from different industrial sectors differently, the principal contemporary example being the corporate income tax. Unincorporated business activities are taxed at lower rates than incorporated firms. Another option governments have in setting income tax rates or schedules is to tax comparable agents in different regions differently, usually with the goal of attaining regional economic development objectives.

Excise taxes are taxes on commodities – indirect taxes. A tax authority can choose to tax all commodities at a uniform rate, at different rates, or exempt some commodities from taxation at all. Excise taxes generally are levied on producers

rather than directly on consumers, but frequently the producer is simply a convenient tax-collection agent for the government. Consumers effectively pay most or all of the tax. Customs duties are excise taxes. Sales taxes are paid by consumers on their purchases of goods and services, and the base of a sales tax is generally much broader than is that of a commodity tax.

Wealth taxes can be collected in several different fashions. One is the bequest, or estate, tax, which is levied on the estate rather than the individuals receiving it. Another option is a one-time levy against the value of wealth at some particular time. Property taxes are a special case of the wealth tax, the tax being levied against only a particular form in which wealth is held. So-called “real” property – land and structures – has been a more common base than has been the more mobile and hideable assets such as jewelry and vehicles.

Several industries traditionally have been subject to taxes relevant to their particular production structures and outputs. Taxes on land are common in agriculture, as have been taxes on animals. Taxes on agricultural outputs also have been easy to collect because of the common necessity to assemble a harvested output for further processing, or the equivalent with animal-based agricultural operations. Landings taxes are available to extract revenues from fisheries. Such taxes may have the dual effect (sometimes accounted a benefit) of dampening overfishing from an open-access, common-property resource. With this tax, the government can effectively exert “ownership” of the fish and impose a “price” on them to fishermen who otherwise would have an incentive to “overfish” – that is, catch fish at rates in excess of natural replenishment of the fish stock. Whether the government chooses to levy the tax at an *ad valorem* rate or as a specific tax can have important consequences for both the fishing industry and the resource. Mining commonly is subject to severance taxes, on a portion of gross output or gross revenues (that is, before deduction of expenses from revenues). Additionally, governments may claim royalties on state-owned minerals, on either a net (after expenses) or gross (before expenses) basis, on

either quantity or on a value basis. Gross royalties on the quantity of mineral output restricts production of lower grade ores; while a royalty on value of production avoids this restriction, it still restricts production of higher cost ores, which may or may not be largely coincident with lower grade ores. In forestry, taxes on yields, or stumpage taxes, shorten rotation times, or where rotation is not practiced, encourage cutting of younger, first-growth trees. Taxes on forestry land restrict the amount of land used for forestry, thereby reducing the supply of timber.

6.3.2 *Taxation 2: effects of taxes*

This subsection addresses some general issues in how taxes affect various aspects of resource allocation throughout an economy. The discussions of distortions and deadweight losses are quite general, and the concepts presented in those topics should be widely applicable to the analysis of taxation in the ancient Aegean and Mediterranean civilizations. When we move into the effects of specific types of taxes on specific aspects of resource allocation, inescapably some of the results become more dependent on institutional structures. The archaeological-ancient historical / philological reader might find some parts of those presentations straining his or her demand for applicability. I have three principal justifications for subjecting you to this material.

First, I believe that many of the tax institutions involved in these analyses may actually find correspondences in ancient institutions. I have stopped short of presenting analyses of minuscule subparagraphs of contemporary tax provisions, although I occasionally may have ventured into some gray areas of applicability. It may be believed that ancient taxation systems must have been quite simple, if for no other reason than the limited means for collecting and processing the information to use more intricate tax systems. The discussion of the Pylos Ma tablets as referring to a taxation and redistribution system could disabuse observers of notions about the simplicity of ancient tax systems,

and Postgate's presentation of components of the seventh- and eighth-century Assyrian tax system lends itself to various interpretations of complexity.⁹

Second, I present a number of models of specific problems or parts of problems, involving the effects of taxes on one thing or another. The goal here is to demonstrate from a number of perspectives how economic analysis proceeds in the investigation of particular topics – how different models illuminate different aspects of a complex set of interactions without any of them being “wrong” for not showing what the others show. (There certainly can be “wrong” or incorrect models though.) Third, if you never venture into the zones of knowledge that you don’t “need,” you’ll never know how close to the edge of your knowledge you are when you’re applying what you do know. In other words, if I occasionally discuss something you may find inapplicable to your day-to-day analytical problems in the study of ancient societies and economies, I am not especially apologetic. I suspect that, with some patience, you may find them useful sooner or later.

Distortions

This is an opportune time to remind the reader of the importance of “marginal conditions” in contemporary economic analysis. Thinking back particularly to the first three chapters, the conditions for efficiency in both production and consumption were the equalization of different marginal conditions. Efficiency in production of any particular product requires equalization of the marginal rate of substitution of one input for another along a producer's isoquant and the marginal rate of substitution of the same pair of inputs in the market, the latter of which is equivalent to the relative price of one factor in terms of the other (or simply the ratio of factor prices). We could also call this the equalization of the marginal rate of technical substitution (along the isoquant) and the marginal rate of market substitution (along the relative price line). For the efficient allocation of inputs across the production of different products, the marginal rate of transformation of one good into another

(accomplished by transferring inputs from one production process to another) must equal the marginal rate of substitution between the two products, which can be represented by either the slope of some point on an indifference curve or the slope of a relative product price line. We can represent this with such acronyms as $MRT = MRS$. Efficiency in consumption requires equalization of the marginal rate of substitution between pairs of goods (the ratio of marginal utilities) and their relative prices. Satisfaction of the marginal conditions (which come from the first-order conditions in our maximization and minimization problems) will give a Pareto-optimum allocation of resources.

Taxes change these marginal conditions. In general, with taxation, $MRT \neq MRS$. These inequalities between pairs of marginal conditions that would be equalized in the absence of taxes are called “distortions.” Frequently, you may read that such and such a tax “drives a wedge” between some pair of marginal conditions; it means the same thing. Figure 6.7 illustrates these facts with a transformation curve between goods A and B. In the absence of taxation, consumer preferences determine the relative prices of A and B as price line p_0 , the slope of which is $-p_B/p_A$. Apply an *ad valorem* tax at the rate t to product A to get price line p_1 , which has the slope $-p_B/p_A(1+t)$. Through the original consumption point X, this post-tax price line cuts the transformation curve at point X instead of remaining tangent

to it (think of the angle between the pretax and post-tax price lines at point X as the “wedge” created by the tax). Consumers now face a higher relative price of good A. Producers will still receive the relative prices represented by price line p_0 , but consumers will pay according to price line p_1 . At the relative prices of p_1 , consumers would not want to consume the combination of goods A and B represented by point X, but would want to substitute into good B; at the original indifference level, this would be point X' (price line p'_1 is parallel to p_1), but that point is outside the transformation frontier and so is unattainable. Falling back along a ray from the origin through point X', we arrive at the transformation frontier at point X'', where price line p''_1 cuts the transformation frontier at a tangency with an indifference curve lower than I_0 . You will hear and read a considerable bit about “distortions,” some caused by taxes, some by other things; Figure 6.7 illustrates the issue.

Figure 6.7 also shows two of the three principal types of effect that a tax has: a substitution effect from point X to point X'', and an income effect from X' to X''. What Figure 6.7 does not show is what has been called the financial effect, which is the structural dodges that consumers make to avoid the tax – for example, incorporation to convert labor income to capital gains; providing executives with in-kind remunerations that may not be taxed; internal pricing of goods made and used within the legal boundaries of a firm (“transfer pricing”) so as to minimize the appearance of profits, and downright evasion.

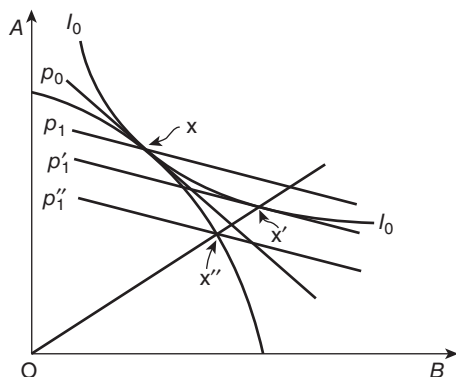


Figure 6.7 Effects of a tax on production.

Deadweight losses from taxation

Several key concepts can be illustrated by developing a simplified, partial-equilibrium view of the effects of a commodity tax. Figure 6.8 shows a demand curve with zero cross-price effects; that is, a change in the relative price of this good induces no substitutions between it and any other goods. It does, of course experience an own-price effect which is a pure income effect. The initial equilibrium, with no tax, settles on price p_0 and quantity q_0 . The imposition of a tax equal to vertical distance $a - a'$ raises the price facing consumers to $p_0 + t$, depresses the price producers

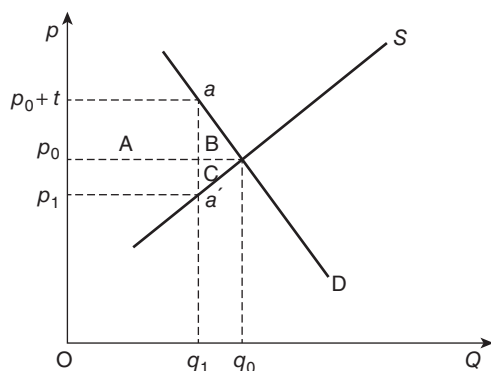


Figure 6.8 Deadweight loss from a tax.

receive to p_1 , and reduces the quantity transacted to q_1 . The tax authority collects tax revenue equal to rectangle A, defined by the points $aa'p_1(p_0 + t)$. The quantity of the product no longer produced or consumed is $q_0 - q_1$. Triangle B, under the demand curve, between quantity q_1 and the demand curve, is the amount that consumers would have been willing to pay for the amount $q_0 - q_1$ but can no longer have; it is the consumer surplus lost because of the tax. Rectangle C is, under certain circumstances, producer's surplus lost similarly. The sum of areas B and C is called the deadweight loss, or excess burden (the part of the burden that is not "excess" being the part consumers and producers pay in tax revenue), of the tax: it is output that simply vanishes because of the inability to attain equalization of marginal conditions between consumers and producers. We assumed away all cross-effects with other goods because the cross-effects could operate in any of three directions: shifting consumers into other taxed goods with greater deadweight losses; shifting them into other taxed goods with lower losses; and shifting them into untaxed goods, reducing government revenue.

Figure 6.9 illustrates one effect that could be visualized in Figure 6.8 by rotating the supply curve clockwise until it is perfectly elastic. With a perfectly elastic supply of the taxed good, the entire burden of the tax payment falls on consumers, because the producers' price is unaffected. There is still a reduction of the quantity

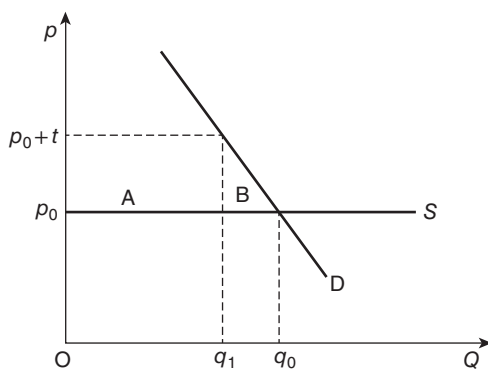


Figure 6.9 Deadweight loss from a tax with perfectly elastic supply.

sold from q_0 to q_1 , but with no change in product price, producers can simply redeploy their resources elsewhere at constant factor prices. Now, imagine rotating the demand curve in Figure 6.8 to be perfectly elastic, leaving the supply curve sloping upward. Under these circumstances, the full amount of the tax is paid by producers. A general formulation for the change in the producer price following the imposition of an excise tax on a single good which has zero cross-effects is $\Delta p / \Delta t = \epsilon_d / \epsilon_s - \epsilon_d$, where ϵ_d and ϵ_s are demand and supply elasticities at the pretax equilibrium.

The question naturally arises, "What tax rate would cause the least deadweight loss?" We have seen in Figures 6.8 and 6.9 some principles that operate to affect both who loses (consumers or producers) and how much, and some generalization can be made on their basis. Suppose we have a number of commodities from which we want to extract some tax revenue. To simplify, assume that none of the commodities have any cross-effects with any others. We set up a minimization problem to minimize the deadweight loss subject to a revenue constraint. Actually, we'll formulate the minimization problem as the maximization of the negative of deadweight loss: $\text{Max}_{t_i} -\sum_i B_i$ subject to $\text{Revenue} \geq \bar{R}$ (Atkinson and Stiglitz 1980, 368–369). The instruments for conducting the maximization are the tax rates, t_i . Satisfying the first-order conditions for

maximization involves adjusting the tax rate for each good to find the rate that minimizes deadweight loss while satisfying the revenue constraint. From some rearrangement of this first-order condition comes a partial-equilibrium rule for “optimal” tax rates: $t_i/(1+t_i) = -k/\varepsilon_{ii}$, in which $k = \lambda/(1+\lambda)$, where λ is the Lagrange multiplier, and the entire term k is an adjustment factor for the overall marginal cost of the public funds. This last formula is known as the “inverse-elasticity rule” for taxation (there are several inverse elasticity rules; think of the monopolist’s optimal mark-up over marginal cost in Chapter 4). According to it, taxes should be highest on goods with low own-price elasticities of demand, what would be called “necessities.” This rule of thumb has been looked down upon by many public finance specialists because of its abstraction from cross-effects and its embodiment of only the efficiency principle of taxation (note that food would be taxed quite heavily under this rule, and such a tax would fall quite heavily upon the poor). Nonetheless, its derivation is close in structure to the way optimal tax systems – Ramsey taxes – are derived, which allow for the full range of substitution, for full utility maximization by consumers, and consideration of the social welfare function which weights the utilities of different individuals, in a unified and comprehensive, utility maximization problem. We will discuss that approach to taxation in a subsequent subsection.

More typical than the introduction of a tax is the change of an existing tax rate, which actually can be expected to impose a larger deadweight loss than the introduction of a tax. Figure 6.10 shows both the introduction of a tax and its increase levied against a good with zero cross-effects. The pretax equilibrium is quantity q_0 and price p_0 . Introduction of a tax $t = a - b$, measured vertically, raises the consumer price to $p_1 + t_1$, depresses the producer price to p_1 , and reduces the equilibrium quantity transacted to q_1 . The government collects tax in the amount equal to the rectangle $(p_1 + t_1)abp_1$, consumers and producers experience the triangles ajc and jcb in deadweight losses. Now, suppose the tax is raised from t_1 to t_2 . Producer price is depressed from

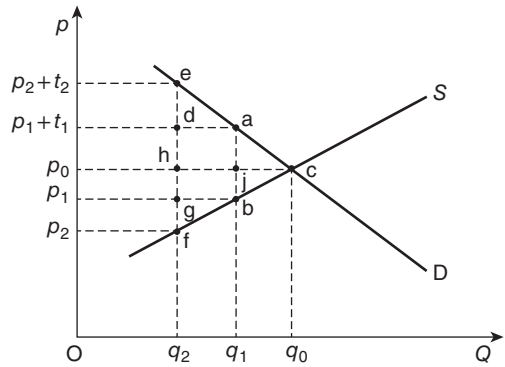


Figure 6.10 Gains and losses to various parties from a tax.

p_1 to p_2 , consumer price rises from $p_1 + t_1$ to $p_2 + t_2$, and the quantity transacted falls further, to q_2 . Government revenue increases by the areas of the two rectangles $p_1gf p_2$ (from producers) and $(p_2 + t_2)ed(p_1 + t_1)$ from consumers, but it decreases by the area of the rectangle $dabg$, the amount it formerly collected on the quantity $q_1 - q_2$, which consumers have stopped buying. This rectangle, plus additional consumer and producer surplus losses of triangles eda and fgb , are added to the initial deadweight loss of triangle abc to form the new, total deadweight loss.

If the change in the tax ($\Delta t = t_2 - t_1$) is small, the additional deadweight loss need not be. Again with no cross-effects, the deadweight loss of a tax increase is approximated by the formula $DW \geq -(\Delta q/\Delta t) \cdot \Delta t[t + \frac{1}{2}\Delta t]$. The last term in this expression shows that the change in the deadweight loss increases as the square of the change in the tax rate. The generality of this result is severely limited by the assumption that there are no cross-effects with the consumption of other goods. When we admit cross-effects into the calculation of deadweight loss, the change can go in either direction. In this more general case, $DW \geq -(\Delta t_i)^2(\Delta q_i/\Delta p_i) - \sum_j t_j(\Delta q_j/\Delta p_i)\Delta t_i$, in which $\Delta p_i = \Delta t_i$. The first term, which represents the reduction in the demand for the good whose tax rate has increased, is always positive. The second term, which is the set of substitution

terms from the Slutsky relationship, represents the changes in demands for complements to and substitutes for goods q_i , goods which themselves are each taxed at some rate t_j . Consumption of complements will fall, with corresponding private deadweight losses and losses of tax revenue to the government. Consumption of substitutes will rise, adding to revenues and consumer and producer surpluses. Without further information, it is not possible to say which effect will dominate, although we can say that if there are pre-existing taxes on substitutes for the good whose tax rate is being increased, the increase in this tax is likely to reduce aggregate deadweight loss throughout the economy. Consequently, a government could be better off with a large number of smaller taxes on a wide array of goods rather than a small number of more substantial taxes on a few. With such a portfolio of taxes, it would have a hedge of sorts against the possibility that the cross-effects term in the last formulation reinforced the direct effect on excess burden whenever it went to increase a tax.

The concepts of the marginal cost of public funds and the efficiency of taxation refer to these deadweight losses. The concepts will appear again in the analysis of optimal taxation and in the evaluation of government expenditures. The magnitude of these losses in contemporary economies is large enough to force serious consideration of them by tax authorities. The French Planning Commission recommended a rule of thumb for the opportunity cost of public funds of 1.5; that is, one franc of tax revenue was expected to displace one-and-a-half francs of private resources, which translates into deadweight losses of 50%, at the margin.¹⁰ Estimates of the marginal welfare costs of raising existing taxes in the United States in 1973 range from 17¢ to 56¢ per dollar of revenue raised, depending on the type of tax and the elasticities of savings and labor supply; the average marginal cost for all taxes combined ranged from 17¢ to 33.2¢ per dollar (Ballard *et al.* 1985, 136, Table 4).

The treatment of these inefficiencies has been in a partial-equilibrium framework so far. We have considered only effects on commodity prices, but taxes frequently are sufficiently

widely applied in an economy to have noticeable effects on factor prices (labor, capital, land) and income distribution as well, which takes us directly to a general equilibrium characterization for a comprehensive – or even satisfactory, sometimes – analysis of the effects of tax changes. We consider these comprehensive effects in the following subsection, but first we turn to the partial-equilibrium analysis of some prominent effects of several types of tax. The partial-equilibrium analysis is nonetheless interesting, if incomplete sometimes, because it highlights some of the basic mechanisms which operate in general equilibrium, if on a broader scale.

Personal income taxes and labor supply

Income taxes in contemporary, industrial societies are particularly complex systems of differential rates and excludable expenses, and it is difficult to obtain clear and comprehensive, analytical results about them.¹¹ However, analysis of some simple situations, gradually building up in complexity to approximate real systems, yields some insights into the mechanisms at work in the more complicated systems.

The substantial majority of income taxed by a personal income tax is labor income. Accordingly, the greatest potential for that tax to affect resource allocation decisions is in the area of labor supply, since an income tax can be represented effectively as a tax on hours worked. Whether we are dealing with a formal labor market or other arrangements for matching the demand for and supply of labor, many of these effects will be much the same, possibly with some differences in observability. We have not yet studied the supply of labor other than as a particular instance of the supply of any factor of production, which might range from capital and land to timber and seeds.¹² We will examine labor in more detail in a subsequent chapter, but for the present purposes, you may think of the labor supply curve as being derived from a utility function in which generalized consumption (a composite consumption good) enters positively and labor time enters negatively – that is, there is disutility to work, which abstracts from any direct pleasure one might derive from labor and

from on-the-job perks. Consequently, we will work with indifference curves between, alternatively, a composite consumption good and labor and the same composite consumption good and leisure. The two approaches are interchangeable, but it will be instructive to see examples of each.

Whether a proportional income tax on all income, with no exemptions, would increase or decrease the supply of taxable labor cannot be determined without further information on the utility function (Atkinson and Stiglitz 1980, 31–36). Figures 6.11(a) and 6.11(b) portray income on the vertical axis and leisure on the horizontal axis, with the all-leisure (no-work) position identified as $\bar{L}$. The upward-sloping lines

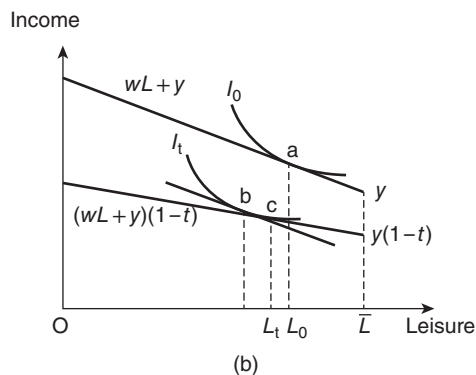
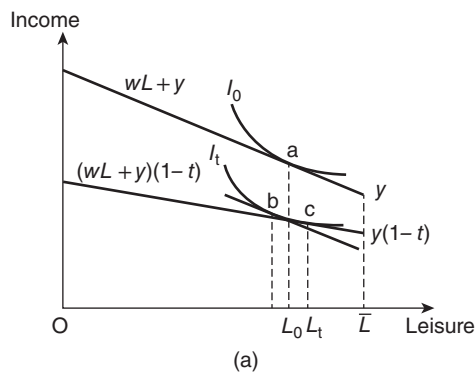


Figure 6.11 (a) Effect of an income tax of labor supply, substitution effect > income effect. (b) Effect of an income tax of labor supply, income effect > substitution effect.

are budget lines that begin at the right-hand side of the diagram, at $\bar{L}$ at a height equal to nonlabor income y . The slope of the pretax budget line is equal to the wage rate. Given her preferences between leisure and income, the consumer with indifference curve I_0 reaches a tangency on the budget line at point a, which gives a pretax labor supply of L_0 . Introduction of the tax has two effects on income, which can be separated according to the Slutsky relationship, $(\Delta L / \Delta t) = (\Delta L / \Delta w)_u \cdot (\Delta w / \Delta t) + L(\Delta L / \Delta y) + (\Delta L / \Delta y) \cdot (\Delta y / \Delta t)$, where the subscripted “u” on the $\Delta L / \Delta w$ term indicates that this change is made holding utility constant. The nonlabor component drops from y to $y(1 - t)$, and the slope of the budget line, which now represents the after-tax wage rate, becomes flatter. The highest indifference level the consumer can reach after the tax is I_t . The pure income effect is the movement from point a, on the pretax budget line to point b, and the substitution effect takes the consumer from point b to point c, for the full, post-tax equilibrium. In both figures, the pure income effect would cause the consumer to reallocate hours from leisure to labor: leisure is a normal good, and income has fallen. In Figure 6.11(a), the substitution effect outweighs the income effect, and in the new equilibrium, the consumer works fewer hours and takes more leisure. In Figure 6.11(b), the substitution effect still works in the opposite direction to the income effect but is not strong enough to outweigh it, and our consumer works more in response to the income tax. The income effect will outweigh the substitution effect in the full response of labor supply to the tax if the marginal utility of income times the share of labor income in total income is less than one. If that product is greater than one, labor supply will decrease.

An income tax has an excess burden just as an excise tax does. Figure 6.12 reproduces the leisure-income basis of labor supply presented in Figures 6.11(a) and (b), but we apply the tax only to labor income for simplicity. Pretax income gives the budget line labeled w , which the tax rotates to $w(1 - t)$. The pretax labor supply, L_0 , is determined by the tangency of the indifference

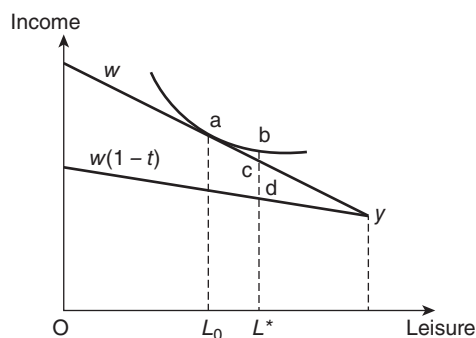


Figure 6.12 Excess burden of an income tax.

curve with budget line w at point a . To determine excess burden, let's keep the consumer on the original indifference level at the new wage rate; move the new budget line $w(1-t)$ parallel upwards to a tangency with the original indifference curve at point b . The movement to point b is a pure substitution effect. If the tax authority redistributes the tax revenue in lump-sum fashion to taxpayers to try to restore them to their original utility level after the tax, it would have collected an amount of tax from a tangency at point b equal to the difference between the original and post-tax budget lines: distance $c-d$ on the diagram, since labor supply associated with tangency b is L^* . However, an amount of redistribution equal to distance bd is needed to fully restore the consumer to his original utility level. The difference between what is required for full compensation and what the tax will collect, bc , is a measure of the excess burden of the income tax.

The analysis of the introduction of, or change in, an income tax is much more complicated when there are multiple tax rates, such as is the case with a progressive tax rate, that is, increasing average tax rates at higher levels of income. Suppose we have a progressive tax schedule with three rates, $t_1 < t_2 < t_3$, which apply to income levels, $m_1 < m_2 < m_3$, where total income $m = wL + y$, with y being nonlabor income again. Figure 6.13 depicts the three rates as line segments of decreasing slope as more labor hours are supplied. The intercept with the vertical axis shows nonlabor

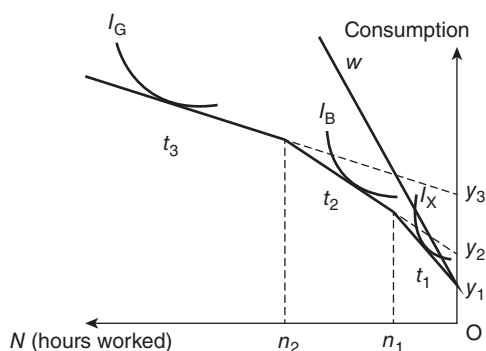


Figure 6.13 Effects on labor supply of a progressive income tax.

income of y_1 . At a constant pretax wage rate, w , indicated by the straight line w , across all hours worked, the lowest tax rate applies to individuals working from zero to n_1 hours, which gives a maximum income of $y_1 + wn_1$. Some individuals, represented by the indifference curve of individual X , will choose to work in this range of hours. Individuals supplying labor hours between n_1 and n_2 will fall into the tax bracket of t_2 ; those supplying more hours than n_2 will be taxed at rate t_3 . The effective wage rate in the first tax bracket is $w(1-t_1)$, that in the second bracket is $w(1-t_2)$, and that in the third is $w(1-t_3)$.

The nonlinear tax schedule creates an oddity in the effect of income on labor supply, shown by the dashed lines in Figure 6.13 extended from the line segments of tax rates t_2 and t_3 back to the vertical axis, and labeled y_2 and y_3 , and called "virtual income" (Killingsworth 1983, 89–91, 333–334, 337–339; Hausman 1985, 217–219).¹³ If an individual in tax bracket 2, that is, working between n_1 and n_2 hours, were given nonlabor income in any amount between y_1 , which he in fact has, and y_2 , the labor supply determined by the tangency of his indifference curve and the tax-rate-2 budget line would not change. The relationship between his labor hours supplied and consumption behaves as if he had property income of y_2 . Thus, the income and substitution effects associated with a property income of y_1 do not operate along the second and third segments of the budget line.

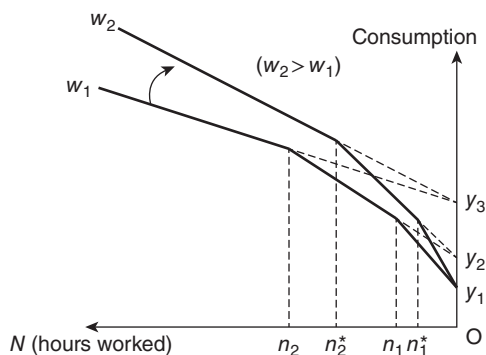


Figure 6.14 Wage increase under a progressive income tax.

An increase in the wage rate, given this tax structure, would rotate the entire, kinked budget line clockwise around y_1 , but would leave the projections of the t_2 and t_3 segments of the budget line onto the vertical axis unchanged at y_2 and y_3 . With the incomes at which tax rates change from t_i to t_{i+1} , the earnings limits of the tax brackets, n_1 and n_2 , would shift toward the origin to n_1^* and n_2^* . The ordinary income and substitution effects operate as long as hours worked remains in the same segment of the budget line. Figure 6.14 shows these movements. The behavior of labor hours depends on the effect of changing virtual income as well as the changing wage rate.

Several separate tax policies are available with this systemic structure. First, different rates could be applied to nonlabor and labor income. Second, the tax rate in each bracket could be changed independently of the rates in the other brackets. Third, the break-points of the tax rate could change (n_1 and n_2 hours, which given the constant wage rate assumed, translate into progressively higher incomes). A change in only the highest tax rate will affect only persons supplying hours within that range of hours; thus an increase in t_3 would rotate the leftmost segment of the budget line counterclockwise, raising virtual income y_3 . The substitution effect would reduce hours, and it could reduce them as far as the lower earnings boundary of the tax bracket, but not beyond and into the lower bracket. A

reduction of the middle rate, t_2 , could, however, move labor hours supplied past the upper edge of that earnings bracket and into rate t_3 . A change in the lowest rate, t_1 , would cause changes in the virtual incomes of all higher tax brackets but no changes in the slopes of the other segments of the budget constraint. A change in one of the earnings limits, n_i , changes all the virtual incomes on subsequent segments of the budget constraint in the same direction. The responses of hours of labor supplied in response to these changes in the tax schedule generally can go in either direction, and further information about the tax code and the utility function is needed to make predictions. Nonetheless, the forces operating can be clearly identified and their partial effects predicted.

If we studied the effects of income taxes in a household production setting, the location of production – within or outside the household – could be affected by the tax structure. The income tax can be thought of as a tax on the production of market-produced goods. As such, it will affect the relative production advantages of various family members and will, in general, divert activity from taxed to untaxed activities.

Effects on saving

In general, saving is equivalent to future consumption. Savings can take a number of forms. For instance, the purchase of a durable good is an act of saving. Construction or modification of a house is a type of savings. Hoarding of money and purchases of financial assets are methods of savings. Investment in oneself, say through education or training, is a type of saving. Most of the determinants of savings fall into one of three major motivations: consumption-smoothing (sometimes called the lifecycle motive), precaution or insurance, and bequests. There are many opportunities to insure many different assets – income, postretirement income, health, houses, vehicles – and much of the precautionary motivation for savings can be studied quite directly by looking at these individual markets. Bequests can be studied as an instance of utility-maximizing altruism, but in practice they frequently contain a large accidental component.

Consequently we will focus our examination of the effect of taxation on savings on the life-cycle motivation.

Many different taxes can affect savings. Quite direct are taxes on interest, taxes on capital gains, wealth taxes, inheritance taxes, taxes on real property, and taxes on corporate profits. Personal and business income taxes contain several of these distinct types of tax, and even a tax on labor income will have effects on savings by virtue of the link between consumption and savings: it will depress saving through an income effect, without changing the relative prices of current and future consumption. A tax on consumption, in contrast, will increase savings during the working period of a consumer's life as a means of deferring the taxes; this takes no extraordinary foresight on the typically myopic consumer's part, just the elevation of the price of current consumption.

Frequently, taxes on capital reduce the after-tax return to capital, making future consumption more expensive relative to current consumption, which tends to increase current consumption. Taxes on capital affect borrowers and lenders differently. The reduction in the post-tax return to capital confers a positive wealth effect on borrowers, raising their present consumption: income and substitution effects in response to a tax on capital operate in the same direction for borrowers. For lenders, the consequence is more ambiguous, depending on the relative magnitudes of income and substitution effects.

We can show some of the mechanisms involved in the taxation-savings relationship. The simplest model of the lifecycle savings mechanism is due to Irving Fisher (1930).¹⁴ In the simplest version, consider a consumer whose utility function contains as its only elements consumption in a first period and consumption in a second period: $u = u(c_1, c_2)$. In the first period, her budget constraint says that her consumption plus her savings add up to her total income: $c_1 + s = y_1$; in the second period her consumption is equal to her income plus whatever she saved in the first period, plus the interest on that saving: $c_2 = y_2 + (1 + r)s$. We can combine these two budget constraints as $c_1 + c_2/(1 + r) = y_1 + y_2/(1 + r)$. Now, let her

maximize this utility function subject to this combined budget constraint by adjusting her first and second period consumption levels: $\mathcal{L} = u(c_1, c_2) - \lambda[c_1 + c_2/(1 + r) - y_1 - y_2/(1 + r)]$. The first-order conditions yield the relationships that $(\Delta u/\Delta c_2)/(\Delta u/\Delta c_1) = 1/(1 + r)$, or alternatively, $(\Delta u/\Delta c_1)/(\Delta u/\Delta c_2) - 1 = r$. The first expression says that the marginal rate of substitution between present and future consumption should equal the price of future consumption in terms of present consumption, which in turn, is just the definition of the discount factor for comparing future values with present ones. The price of future consumption in terms of current consumption is $1/(1 + r)$; if the interest rate rises, this value gets smaller, which means of course that the relative price of future consumption has fallen; deferred present consumption "grows" at a faster rate, thus yielding a larger amount to be consumed if left to grow for a period. The second version of the expression says that the marginal rate of time preference should equal the interest rate.

It is useful to look at the Slutsky equation for this model to see how the interest rate affects present consumption. Forming the demand function for first-period consumption as $c_1 = f(r, y_1, y_2)$ and performing a number of manipulations on it that involve variables from both periods, we can arrive finally at $\Delta c_1/\Delta r = (y_1 - c_1) \cdot (\Delta c_1/\Delta y_2) + (\Delta c_1/\Delta r)_{u^*}$. The subscript " u^* " in the second term (the substitution effect) indicates that the change in c_1 caused by a change in r is a movement around a constant level of utility; this term is negative. It is obvious that the coefficient of the first term, which is equivalent to savings (s), can be negative or positive. For borrowers, who consume more in period 1 than they bring in income, it is negative, and both the income and substitution effects are negative: a higher interest rate depresses current consumption and correspondingly raises saving. (Think of how you use credit cards as interest rates fluctuate.) For lenders, this coefficient is positive, and the net effect of a change in the interest rate for them is an empirical matter.¹⁵

Using $1/(1 + r)$ as the price of second-period consumption, the elasticity of first-period

consumption with respect to that price is equal to the savings rate times the difference between the elasticity of substitution between first- and second-period consumption and the wealth elasticity of first-period consumption. (Wealth is the present discounted value of both periods' incomes, including any asset income.) With a tax on savings, the price of second-period consumption becomes $1/[1 + (1-t)r]$. Whether saving rises or falls in response to an increase in the tax rate depends on an intricate set of relations between the savings rate, the elasticity of substitution between consumption in the two periods, the interest rate, and the wealth elasticity of first-period consumption.¹⁶ The larger is the elasticity of substitution between consumption in periods 1 and 2, the larger is the wealth elasticity of first-period consumption required to cause savings to rise in response to a tax increase.

A tax on interest income will reduce after-tax r , raising consumption for borrowers unambiguously. Consequently, the interest elasticities of consumption and saving are of particular interest to tax policy. An estimate of the interest elasticity of savings for the United States in the 1970s was around 0.4, implying a consumption elasticity around -0.3. Being for a particular time and place, these estimates depend on the corresponding configuration of preference and wellbeing parameters and institutional structures, but the sizeable magnitude is worth remembering nonetheless.

We can examine the effects of different types of tax on savings by inserting those taxes into the budget constraint. An equal proportional tax rate, t , on both labor and capital (interest) income is given by: $c_1 + c_2/[1 + r(1-t)] = y_1(1-t) + y_2(1-y)/[1 + r(1-t)]$. The tax has the effect of reducing the effective interest rate, so the substitution effect encourages present consumption. The income effect strikes present and future income equally. Indirect taxation places the excise tax, t_e , on consumption rather than on the sources of income; this is represented by: $(1 + t_e)c_1 + (1 + t_e)c_2/(1 + r) = y_1 + y_2/(1 + r)$. Recall that we observed early in this chapter the equivalence between a uniform tax on all

goods and a uniform tax on all income. We can re-express the budget constraint for the excise tax to show a partial version of this equivalence: $c_1 + c_2/(1 + r) = y_1/(1 + t_e) + y_2/[(1 + t_e)(1 + r)]$. The excise tax is equivalent to a tax on labor income and has no differential effect on the time profile of consumption. Any effect on savings will be a pure income effect. Different excise tax rates for the different periods certainly would interpose such a relative price change.

The two-period model is restrictive in the sense that it constrains consumption in the two periods to be substitutes. The substitution effect of an interest rate change on future consumption is positive (interest rate goes up, consumption tilts toward the future; if r goes down, consumption tilts toward the present). With more than two periods, consumption in any future period can be either a substitute for or a complement to present consumption, and the interest rate effect on present consumption could be positive.

We can extend the Fisher two-period model by incorporating labor supply decisions into the utility function that yields the savings decision. Let the utility function be $u = u(c_1, c_2, L)$, in which L is leisure in the first period and $L + h = T$, where h is hours of work and T is total time. The consumer works in the first period and lives off his savings in the second. The budget constraint is $c_1 + c_2/(1 + r) = w(T - L) + A$, where A is asset income. This model has three commodities (consumption in periods 1 and 2, and labor/leisure in period 1) and two relative prices, w and $1/(1 + r)$ (or just r). The first-order conditions from this problem are that $(\Delta u / \Delta c_2) / (\Delta u / \Delta c_1) = 1/(1 + r)$ and $(\Delta u / \Delta L) / (\Delta u / \Delta u_1) = w$. With positive income effects and negative substitution effects, we find that $(\Delta c_2 / \Delta r)_{u^*} > 0$ and $(\Delta c_2 / \Delta w)_{u^*} < 0$: an increase in the interest rate raises future consumption and a higher wage rate reduces leisure. Using these relationships, at a constant level of utility, the effect of a higher interest rate on first-period consumption, and hence on saving, is ambiguous, despite the unambiguous effect on second-period consumption: $(\Delta c_1 / \Delta r)_{u^*} =$

$-w(\Delta L/\Delta r)_{u^*} - [1/(1+r)](\Delta c_2/\Delta r)_{u^*}$ can be positive or negative. Consequently, a tax that depressed the interest rate would have an ambiguous effect on savings; whether it increased or depressed savings would be an empirical matter.

So far we have treated the interest rate and future income as certain, but those two variables are notoriously uncertain. When we allow for this stochastic character of either of these variables, taxation affects not only their expected values but their variances and higher moments of their probability distributions.¹⁷ Allowing for this uncertainty can change some of the predictions yielded by the certainty models we've used so far. For example, the two-period model concludes that the substitution effect on present consumption of a tax on interest income is positive, since the tax is equivalent to a reduction in the interest rate. But if the interest rate is a stochastic variable, the introduction of a tax, or an increase in an existing one, reduces the variance of the interest rate as well as its expected value. The substitution effect of reduced riskiness is likely negative, leaving the overall substitution effect of the tax ambiguous.

Effects on risk taking

Loss offset provisions in a tax code provide a powerful instrument for government to share in the riskiness of private investments. With full loss offset, the full amount of capital losses is deductible from tax obligations. With partial loss offset, the taxpayer may deduct some fraction less than one of his losses from tax obligations. The specifications of taxes we have used so far imply full loss offsets, but we turn our attention here to the consequences of these provisions in the case of risky assets, which we just barely touched upon with the stochastic interest rate and income just above. Government can take on greater social risk (contrasted with private risk) by accepting the more variable stream of tax revenues that accompanies loss offset provisions.

We'll develop the treatment of risk-taking with a series of models that deal with progressively more general situations.¹⁸ Our formal treatment of saving in the previous subsection did not admit risk into the properties of the single

capital asset. In the first model below we study the influence of taxation on the allocation of wealth between a riskless asset and a single risky asset. We follow that with a model with two risky assets, in which the risk between the two assets may be correlated. Finally we introduce multiple risky assets, called a portfolio, although we will not develop that analysis as fully as the results from it are much less conclusive than in the simpler cases.

In the first model, the individual's utility-maximization goal involves choosing the allocation of his initial-period wealth so as to maximize the *expected value* of his terminal-period wealth. Utility is a function of wealth, W : $u = u(W)$. The consumer can allocate his wealth in the initial period between an asset, M , which has no return (it does not bear interest, but neither does it lose value through price-level change) but is riskless, and an asset A with a stochastic (risky) rate of return of $\tilde{r}$ (a tilde over a variable indicates that it is a stochastic, or random, variable). His initial-period budget (wealth) constraint is $W_0 = M + A$. At the terminal period, his wealth is $W_T = M + A[1 + \tilde{r}(1-t)]$, or $W_0 + A\tilde{r}(1-t)$. We can substitute the terminal-period budget constraint into the utility function and take the expected value of that utility (a procedure we will not show explicitly) by, more or less, adding up products of all the possible values times their probabilities of occurrence. While the first-order condition is important, it is not particularly informative for our purposes; from the second-order condition for utility maximization can be derived the result that the elasticity of A , the allocation of initial-period wealth to the risky asset, with respect to the tax rate, (ϵ_t^A) is equal to $t/(1-t)$.

The first striking aspect of this result is that the response is positive: the introduction of a tax, or the increase in the rate of an existing one, will cause investors / consumers to shift their wealth into the risky asset. Following this rule allows the investor to keep the probability distribution of final expected wealth the same as the initial distribution he created with his allocation of wealth between money and the return-bearing asset. That distribution was

optimal in the initial period, and nothing has been introduced that would cause the optimal distribution to change in the terminal period. The second striking aspect is that the result does not depend on the investor's attitude toward risk – it doesn't matter whether he is a risk-avertor or a "plunger."

The first complication of this simple model is to replace the riskless, zero-return money with a second risky asset with rate of return $\tilde{r}_1$. Going through a similar maximization procedure yields a tax-rate elasticity of the initial-period allocation of wealth to the riskier of the two assets of $\varepsilon_t^A = 1/(1-t) - \tilde{r}_1 t \varepsilon_{tW_0}^A / [1 + (1-t)\tilde{r}_1]$, in which $\varepsilon_{W_0}^A$ is the initial-period-wealth elasticity of the riskier asset (that is, the percentage change in the amount of the riskier asset held, divided by a 1% change in total, initial-period wealth, which induced the change in the holding of the riskier asset). The elasticity $\varepsilon_{W_0}^A$ is positive (it does not say that the *percentage* of the portfolio held in the riskier asset will increase with an increase in total wealth, only that it will not decrease). In this case, the availability of the less risky asset depresses the otherwise positive effect (which is the same as in the case of the single risky asset), but with reasonable values for the tax and safer rate of return parameters (say, $t = 1/2$ and $r_1 = 0.05$), the wealth elasticity would have to be greater than 41 (an enormous number! – a factor of 41 times the size of the event that precipitated the response) to give a net negative ε_t^A .¹⁹

Moving on to the case of multiple risky assets with correlated returns introduces some serious complications for the concept of subsidizing risk. The concept of correlated rates of return bears some explanation. As long as rates of return are stochastic, we can only offer predictions of what they will be in future periods. But, by recording how the rates of return on different assets move over time, we can identify assets whose rates of return tend to move up and down together and those whose rates of return tend to move oppositely – when one goes up, the other tends to go down. The co-movements are called correlations, or covariances (correlations not divided by the product of the standard deviations), between

the rates of return of these assets. With correlated rates of return, it is difficult to distinguish clearly between the riskiness of the assets, as each asset can either contribute to or dampen the riskiness inherent in the other. Consequently we cannot identify one of the assets as the riskier and examine how taxation affects its weight in the portfolio. We can, however, derive the same kind of elasticity for each asset separately. The tax rate used here, t_i , is defined as the tax on the differential return of the i^{th} risky asset over some numeraire asset. The own-asset effect is $\varepsilon_{t_i}^{A_i} = t_i/(1-t_i)$, and the cross-asset effect is $\varepsilon_{t_j}^{A_i} = 0$, when subscripts i and j refer to different assets; that is, the tax rate on one asset (j) does not affect the choice of another asset (i).

In a full portfolio framework, with many risky assets, there are few general results beyond these just presented. Whether government can, on balance, share risk taking through the tax system becomes open to question, and the issue is more doubtful the more private (market) opportunities there are for risk sharing (that is, the more developed are various insurance instruments).

6.3.3 Taxation 3: tax incidence (who really pays?)

The subject of tax incidence is, quite directly, "Who pays the tax?" It is simplistic to believe that whoever has the statutory obligation to pay the tax is the individual who actually pays it. Sometimes a statutorily designated tax payer is little more than a convenient revenue collector for the tax authority. The method by which statutory tax liability is pushed along the transactions channels via a series of price adjustments is called "shifting" of the tax burden. It is possible that more than 100% of a tax burden would be shifted from one group of statutorily liable individuals to some other group or groups. Shifting can occur between groups of factor owners (say, between capital owners and workers – a case that is complicated in practice when workers also own capital), between residents of different regions or countries, between different segments of the personal income distribution (say, from rich to

poor or vice versa), between generations, and probably between other pairs of categories we have not thought of.

When firms are taxed (say, with an income tax), they can raise their product prices or lower their factor payments by direct price reductions where that is possible or by substituting from one factor to others. When the firms are price takers in their factor markets (as is the case with most firms most of the time), they cannot unilaterally pay their factors less than they could obtain in other employment, so factor substitution is the principle method of shifting. We can think of shifting in the case of tax exemptions also: who “pays” for a tax exemption? An investment tax credit in specific industries will divert investment from other industries not receiving the credit to the industry or industries receiving it. The expansion of investment in the receiving industries may lower the prices of their products, while in the industries from which the investment is diverted, dampened supply capacity will tend to raise product prices. The extent to which the industries qualifying for the investment tax credit actually benefit from it depends on the supply effects of their own investment expansion and demand elasticities for their products.

For consumers faced with a tax or tax increase on a consumption item (say, an excise tax), their primary method of shifting is substitution into other products. This may raise the price of substitutes and lower the price of the taxed commodity, partially shifting the burden of the tax payment onto other consumers of the substitute commodity.

Whereas most of our assessment of taxation so far has been in partial-equilibrium terms, tax incidence generally is a general-equilibrium matter. When shifting occurs through changes in factor prices, which it usually does, we are in a realm in which the functional distribution of income (that is, by category of factor) changes and demands for products can change, further affecting factor prices. The imposition of, or changes in, very specific taxes may or may not have sufficiently broad ripples through an economy to justify a general equilibrium analysis. Some judgment is involved in deciding when

a partial-equilibrium incidence assessment is satisfactory.

The parameters that are important in determining the shifting, and hence the incidence, of a tax are product demand elasticities, both product and factor substitution elasticities, factor input ratios in different industries (for example, capital/labor ratios), and factor supply elasticities. Other critical issues in determining incidence are the specificity of the tax (by industry, region, factor, product, and so on), the mobility of factors between taxed and untaxed (or generally between differentially taxed) sectors or regions, and which prices, if any, are fixed (that is, they can’t change).

The best known and most widely used general-equilibrium model of tax incidence is the Harberger model of the corporation income tax (Harberger 1962; McLure 1975). While the model is named after the particular tax Harberger studied, the model itself is a completely general framework for studying the incidence of sector-specific taxes under a wide range of conditions of intersectoral factor mobility. To give a flavor of the relationships involved in modeling tax incidence in general equilibrium, we describe (but do not show) the components of the Harberger model. First, there is a specification of the demand for the output of the taxed sector (the comparable relationship for the untaxed sector is unnecessary). Second is the production function for the taxed industry (again, the comparable function for the untaxed sector is unnecessary because any factors not employed in the taxed sector must go into the untaxed sector by virtue of the full-employment assumption). Third are the relationships specifying the substitutability between factors in each industry (specifications for both industries are needed for this relationship: just because we know that the other industry has to soak up any units of factors released by the taxed sector doesn’t imply that we know how “easily” it does so). Fourth is a set of formulas relating gross-of-tax and net-of-tax factor prices and the taxes themselves. Fifth is a set of factor supply specifications. Sixth is a set of relationships describing intersectoral factor mobility in terms of either factor prices or quantities. The last relationship needed to

“close” the model (that is, make the number of independent variables equal the number of separate relationships) either specifies a numeraire price or the description of a particular tax policy.

We offer an intuitive walkthrough of the imposition of a tax in the Harberger model, with an example adapted from Atkinson and Stiglitz 1980, 164–165, 175). Begin with the levying of a tax on income in the corporate sector. The other sector is just the unincorporated sector. Any division of the industries in an economy into “protected” and “unprotected,” taxed and untaxed, or generally taxed differentially on whatever basis would be equivalent to the corporate-noncorporate distinction. We don’t need incorporation laws and a stock market to find the model applicable. The cost of capital rises from r to r times some function of the tax rate. This shifts the supply curve of the taxed sector upward, causing the price of its output to rise. This price increase of its output drives up the demand for the output of the untaxed sector, depending of course on the degree of product substitutability in demand. Two effects follow from this point: a reduction in the demand for capital and a compensating increase in the demand for labor, *per unit of output*, in the taxed sector; and a change in factor demands resulting from a shift of production from the taxed to the untaxed sector. The second of these effects could either exacerbate the reduction in the demand for capital or dampen – even reverse – it, depending on the relative capital intensities of the two sectors (the simplest version of the model uses only two factors – capital and labor). If the taxed sector is the more capital-intensive of the two sectors, it will release more capital per unit of reduced output than the nontaxed sector can use per unit of its expanding output, and the demand for capital drops further (you can see the Rybczynski effect operating here). Whatever the consequent changes in factor demands are, they are equilibrated by factor-price changes. These changes in factor prices precipitate further changes in factor intensities in both sectors, determined by each sector’s production function, and they may lead to a change in the pattern of aggregate

demand if the different factor owners have different preferences (or even different per capita incomes). The adjustments described above have operated through factor substitution and output adjustments. If there is a zero elasticity of substitution in either industry’s production function, the full adjustment will have to take place through output adjustments. If the taxed sector is relatively capital intensive, the ratio of the rental on capital to the wage rate will fall (as we described above). Alternatively, if the taxed sector were relatively labor intensive and had a zero elasticity of substitution in production, it would release more labor than capital, and the expanding untaxed sector would become more labor-intensive as it expanded its output, reducing the marginal productivity of labor relative to that of capital; the rental on capital would rise relative to the wage rate. Even with a positive but small elasticity of substitution in a labor-intensive taxed sector, the wage/rental ratio could fall if the demand elasticity of the taxed sector’s output were small enough or if there were an especially large difference in factor intensities between the taxed and untaxed sectors.

The example we just considered was the incidence of a sector-specific tax on the income from one factor. We turn to a consideration of the incidence of excise taxes. We distinguish between consumer prices, p^c , which include the tax, t , and producer prices, p^p , which do not and are also determined by the cost function, $c(p^p, w)$, where w represents factor prices. We allow for substitution effects among commodities in final demand. The effect of the taxes on the entire range of consumer prices is as follows: $(\Delta p_j^c / \Delta t_i) = \delta_{ij} + (\Delta p_j^p / \Delta t_i)$, in which $\delta_{ij} = 1$ if $i = j$ and zero otherwise, representing the direct effect of the tax on the consumer price. The $(\Delta p_j^p / \Delta t_i)$ term represents the own-price effect of the tax on the producer price if $i = j$ and the cross-price effect if the good under consideration is not the one to which the tax change applies. Thus, if the tax on good i changes the producer price of good i , then the total change in the consumer price is the one-for-one change caused by the tax change itself plus the secondary effect on the producer price. Accordingly, in Figure 6.8,

the producer price being depressed by the tax, the consumer price would not rise by the full amount of the tax; in Figure 6.9, the producer price was fixed (we'll discuss below what "fixed" both means and implies), so the entire amount of the tax went into the consumer price. But what causes changes in producer prices?

Let's turn to the effect of a set of tax changes on producer prices. The formulation of this set of changes is $\Delta p_j^p / \Delta t_i = \sum_k [(\Delta c_j / \Delta p_k^p) \cdot (\Delta p_k^p / \Delta t_i)] + \sum_m [(\Delta c_j / \Delta w_m) \cdot (\Delta w_m / \Delta t_i)]$.²⁰ The subscripting bears some explanation. Once again, in considering the effects of changes in the entire set of excise taxes at one time, we use subscript j for the particular good (whether we're looking at the consumer price or the producer price) and subscript i for the tax to give us a compact expression for referring to cross-effects of taxation when i and j don't refer to the same product. When $i = j$, we're talking about own-effects. You also could think of many sets of rows of the same formulas, with each set containing the entire array of taxed (and untaxed) goods and the rows of each set referring to the effect of the tax on just one of those goods; the next set of rows would contain one row each for the entire array of commodities, but with the effects of the tax on another good; and so on.²¹ Note that the subscript on the cost, c_j , matches the good identified by the producer price; that's because we're considering the effects on the production cost of the good whose producer price we're studying; makes sense so far. Now, for the subscript "k" on the next producer price: some goods may be used as inputs for other products, but those goods aren't necessarily – and probably aren't – the same as the one being produced, which is referenced with the subscript j . This first term is added over all the k tax effects on the producer prices of all the other goods used in making good j . The second term is summed over "m" because we're identifying the factors (labor, capital, and land – the "original" factors as contrasted to the produced ones, sometimes called intermediate inputs or intermediate products or producer goods) used to produce good j . Look back to the cost function in the previous paragraph and you'll see that both intermediate goods and factors are identified by

their prices (the p^p and the w) as being in the cost function.

Notice from the formula for the effect of the taxes on the producer prices that we have $\Delta p^p / \Delta t$ terms on both the left-hand and right-hand sides of the expression, even though the set on the right-hand side, for any good j , contains more commodities; never mind that – a matrix formulation can handle that. What that means for our ability to formulate the incidence issue is that we can re-write that formula as a combination of just $\Delta p^p / \Delta t$ and $\Delta w / \Delta t$ terms, with coefficients (that is, things multiplied by those terms) that contain the information from the cost function on the effects of the intermediate good and factor prices on costs: $\Delta p^p / \Delta t = (\Delta p^p / \Delta t) \cdot A + (\Delta w / \Delta t) \cdot B$. (If it seems odd that we've written the A and B coefficients after the $\Delta p^p / \Delta t$ and $\Delta w / \Delta t$ terms, that's because we're working with matrices now; each of these terms is a big set of columns, with each row referring to a particular commodity and each column doing the same; the number of rows equals the number of columns in the case of $\Delta p^p / \Delta t$; in the case of $\Delta w / \Delta t$, the number of rows equals the number of commodities produced and the number of columns equals the number of factors. This also makes for what may look to some readers as funny methods of multiplication and division, but that's because matrix multiplication and division are more intricate than the same operations on single numbers – "scalars.") We can rearrange the last formulation as $\Delta p^p / \Delta t = (\Delta w / \Delta t) \cdot B(I - A)^{-1}$, in which I is the "identity matrix," or a square matrix (number of rows equals number of columns) with 1s in the "diagonal," or in the cells of the matrix for which $i = j$ (the row and column have the same number, numbering from the upper left-hand corner; the "diagonal" goes from the top left corner of a square matrix to the lower right corner). The scalar (single-number) equivalent of the $I - A$ term is just $1 - a$, where a is a cost coefficient smaller than 1, only in matrix form it represents a whole lot of $(1 - a)$ s, one for each cost function. The exponent -1 indicates that we're essentially dividing by the matrix $I - A$, in matrix terminology, multiplying by the inverse, which amounts

to the same thing; again, with scalars, this would look like $b/(1-a)$. Returning from the short excursion into matrices, what you should notice from the last formula is that the effects on producer prices – the $\Delta p^p/\Delta t$ – depend only on the effects of the taxes on the factor prices. Unless the taxes affect factor prices, the full amount of the taxes (or tax changes) will be paid by consumers.

Now, let's go back to the expression for the effect of the taxes on the consumer prices. Suppose that, for some reason, the final consumer prices were unable to change. In that case, the full amount of the taxes would have to fall on the producer prices, which would, in turn, cause the necessary changes in the factor prices to accommodate them. Under what circumstances could consumer prices be unable to move? In what is called the “small open economy” case in international economics. “Small” means that the country in question is too small to affect world prices of products (ignoring transportation costs). “Open” means that the economy trades with the rest of the world, not necessarily without trying to fiddle with border prices through any of a series of import and export duties or quotas. So, if we have a small country that trades, the prices of its tradable²² final consumption goods will be determined by international prices. *Ergo*, if it tries to impose excise taxes on its consumption goods, either imported or domestically produced, it will just depress the prices of the original factors used in producing them – labor, and land in particular; capital itself might be internationally mobile, making its price difficult to affect by local policies. Such a country might just as well tax factor incomes, and then just the ones that are not highly mobile internationally. (A large country that trades will have some influence over world commodity prices.) The Phoenician littoral states in the early first millennium B.C.E. were small, open economies. Excises imposed on their consumption goods would have been forced backwards onto labor and land, possibly onto capital to some extent as well. It would be interesting to see what can be learned about their tax policies, given this set of relationships.

In open economies, it makes a difference whether taxes on capital are levied on income received by residents or on capital incomes of firms. The former will act as lump-sum taxes on domestic owners of capital. The latter will be escaped pretty much fully by foreign owners of capital, and will drive away foreign investment and may fall instead, fully on local wages and land rents.

The basic lessons of tax incidence theory are that real and nominal (or, effective and statutory) tax burdens are not necessarily related to each other. Consequently, taxes on capital may be borne by workers instead, investment incentives may injure capital owners, taxation of foreigners may simply be equivalent to taxes on domestic factors, and future generations may support those currently alive.

6.3.4 Taxation 4: optimal tax systems

In minimizing the deadweight losses from a tax in section 6.3.2, we largely ignored the interactions of the single, taxed good and other goods, and in coming up with the inverse elasticity rule we had only an efficiency condition, totally ignoring the distributional consequences of the tax, which we did at least point out were likely to strike poorer individuals more heavily than wealthier ones – not a particularly attractive program to recommend, at least by contemporary social and political standards. The elements of a *system* of taxes that we could call optimal should have the following properties then. It should deliver the target revenue – otherwise, what's the point? Second, it should minimize the deadweight losses associated with collecting taxes. Third, it should let us, as a society, choose how heavily the burden of taxation falls on different segments of our public (citizenry, subjects, whatever).

The first element is not a major problem. At the very worst, we can always keep raising some tax rates, although the efficiency properties of such a solution would not be very good. The second two elements raise the classic tradeoff between equity and efficiency, but the social welfare function is a construct that helps clarify the issues involved in that tradeoff. The methodology of optimizing

a tax system with respect to these three elements is to maximize a social welfare function subject to the government's revenue constraint; the constraint that, individually, consumers maximize their own utilities; and technological feasibility constraints (that is, constraints imposed by production functions). This theory has evolved out of the Ramsey model of optimal taxation, which we present briefly as a precursor (Ramsey 1927).²³

The theory of optimal taxation is a complex and relatively difficult subject, full of complicated, formulaic expressions that are not particularly intuitive. Neither is the subject particularly illuminated by diagrams that can show in broad brush strokes the principal mechanisms at work. I have chosen to subject you to as much of it as I do below, as simplified as I have been able to make it, because of the importance of the subject. The theory does give some insight into why tax authorities may have evolved at least some of the relative tax levies they have (aside from rank political machinations), or at least why some have lasted while others were modified. Additionally, it provides a unification of tax policy and income redistribution, a subject that has been of some centrality in the study of ancient economies and societies throughout the Mediterranean. Some of the concepts from the theory may be of use to analysts trying to discover both motivations and consequences of policies for which at least partial evidence remains.

Begin with the tax authority maximizing the welfare of a representative consumer (we'll expand this to consider different individuals, which gives substance to the issue of income redistribution through the tax system). The representative consumer supplies L units of labor and consumes the quantity x_i of goods i (the numbers we use to describe specific commodities run from 1 through some arbitrary number n , where n is the total number of commodities we're considering); he maximizes a utility function $u(\mathbf{X}, L)$, where $\mathbf{X}$ is a compact notation for all the x_i , subject to his budget constraint, $\sum_i (p_i + t_i)x_i = wL$, where w is the wage rate. (The government could apply the tax on the wage rate if it chose, rather than use indirect taxes, and under some conditions, the two types of tax would be equivalent.) Now, the tax authority maximizes individual welfare for the representative individual subject

to its revenue constraint and the individual's first-order conditions for utility maximization. This can be expressed in terms of indirect utility maximization. Assuming commodity prices to be fixed, we can express the indirect utility function as $v(\mathbf{t}, w)$, in which unsubscripted $\mathbf{t}$ represents an array of taxes t_i . The Lagrangean function then is $\mathcal{L} = v(\mathbf{t}, w) + \lambda(\sum_i t_i x_i - \bar{R})$, where $\bar{R}$ is the target revenue. The tax authority adjusts the tax rates t_i so as to satisfy first-order conditions for each one and for the Lagrange multiplier so as to satisfy the revenue constraint. For each tax rate t_i , the first-order condition is $\Delta v / \Delta t_i = -\lambda(x_i + \sum_j t_j (\Delta x_j / \Delta t_i))$. Rearrangement of this formulation, using some "tricks" from the indirect utility function and the terms in the Slutsky relationship that break the total price-responsiveness of a good into substitution and income effects, we can get a rather complicated expression for the optimum tax on each good: $\sum_i t_i s_{ji} = -[1 - \sum_i t_i (\Delta x_i / \Delta y) - (\alpha / \lambda)]$, in which s_{ji} is the Slutsky substitution term, y is the representative consumer's income (or wL) and α is his marginal utility of income. If there are no income effects ($\Delta x_i / \Delta y = 0$), and the cross-price effects are all zero ($s_{ji} = 0$ for all $j \neq i$), this Ramsey tax simplifies an expression quite close to the partial equilibrium formula developed above: $t_i = -[(\lambda - \alpha) / \lambda] \cdot (x_i / \epsilon_{ii})$, in which the term on the right-hand side has a minus sign because the own-price elasticity of commodity i is negative.

Consider an example with two goods and labor, adapted from Sandmo (1976, 46–47; Atkinson and Stiglitz 1980, 375–381, 386–390). The rather complicated formula in the previous paragraph takes the explicit form $t_1 s_{11} + t_2 s_{12} = -k x_1$ (where k stands for a long, complicated, and unenlightening coefficient of x_1) and $t_1 s_{21} + t_2 s_{22} = -k x_2$. This is a set of two equations in two unknowns, t_1 and t_2 . We can solve for t_1 and t_2 in terms of x_1, x_2 , the s_{ij} , and k . Working from the relationships between own- and cross-price elasticities of demand and the s_{ij} terms, we eventually come up with the relationship between the two taxes that $t_1 / (1 + t_1) = [t_2 / (1 + t_2)] \cdot [(\epsilon_{12} - \epsilon_{22}) / (\epsilon_{21} - \epsilon_{11})]$, in which the ϵ_{ij} are own- and cross-price elasticities of compensated (Hicksian) demand for goods 1 and 2. Thus the relationship between optimal tax rates is a factor of a ratio of own- and cross-price elasticities of demands for both taxed goods,

not simply the own-price elasticity of one of them alone.

We can extend this example to consider the relationship between the purchased commodities and leisure and the implications for optimal taxation. We introduce the cross-price elasticities between leisure and the first and second commodities of the previous paragraph, ε_{1L} and ε_{2L} . From the properties of own- and cross-price elasticities of demand for any good, we know that $\varepsilon_{ii} + \varepsilon_{ij} + \varepsilon_{iL} = 0$ for both goods. Substituting into the optimal tax rule for goods 1 and 2, we get the new rule that $t_1/(1+t_1) = [t_2/(1+t_2)] \cdot [-(\varepsilon_{11} + \varepsilon_{22}) - \varepsilon_{1L}]/[-(\varepsilon_{11} - \varepsilon_{22}) - \varepsilon_{2L}]$, which tells us that the tax rate ought to be higher on the good that has the lower cross-elasticity with leisure. Stated alternatively, goods that are complementary with leisure should be taxed more heavily, because leisure itself cannot be taxed directly.

Now let's advance to the problem of optimal taxation with redistribution. We consider an economy with a large number of individuals who differ in their income levels. Rather than use the technique of the representative consumer, which assumes that all consumers are identical, we have the tax authority (the government) maximize a social welfare function, $W(v_1, v_2, \dots, v_H)$, in which v_h are the indirect utilities of individuals (households) 1 through H . The Lagrangean to be maximized is $\mathcal{L} = W(v(t)) + \lambda[\sum_{i=1}^n t_i(\sum_{h=1}^H x_i^h) - \bar{R}]$. Without showing the first-order condition from adjusting the t_i , rearrangement of it and compacting some notation yields the formulation for the optimal set of commodity taxes: $\sum_i t_i \sum_h s_{ik}^h / H / \bar{x}_k = -[1 - \sum_h (b^h / H)(x_k^h / \bar{x}_k)]$, which requires some explanation of the terms that appear in it. First, H is just the number of individuals or households, so dividing by H gives an economy-wide average. Next, $\bar{x}_k$ is the economy-wide average consumption of good k . The term b^h is more intricate, but it is the entry point for evaluating income transfers. Its formulaic definition is $(\beta^h / \lambda) + \lambda_i t_i (\Delta x_i^h / \Delta y^h)$, in which β^h is the marginal social valuation of income for household or individual i : $(\Delta W / \Delta v_i) \cdot (\Delta v_i / \Delta y_i)$ (we will use below the further shorthand that $(1/H) \sum_h b^h = \bar{b}$). This term b^h is the value of transferring one unit of tax revenue to household h , net of the additional tax paid through spending the transfer on taxed commodities. From this

expression emerge two distributional rules. A tax rate on a commodity is lower: the more the good is consumed by households with a high social marginal valuation of income (that is, low-income people, increments in whose income have a high marginal valuation in the social welfare function – assuming that's the kind of social preference structure we're working with); and the more the good is consumed by households with a high marginal propensity to consume the taxed good.

This rule can be written in an alternative fashion that helps bring out the efficiency-equity tradeoff more directly: $\sum_i t_i \sum_h s_{ik}^h / H = -x_k(1 - \bar{b}r_k)$, in which $r_k = [\sum_h (x_k^h / \bar{x}_k)(b^h / \bar{b})] / H$. This term r_k , minus 1, is the covariance between the consumption of commodity k and the net marginal social valuation of income, or alternatively an indicator of the propensity of low-income households to consume good k relative to the propensity of other income groups. When this term is large, the tax on commodity k should fall. If we eliminated income and cross-price effects to simplify the problem, this rule would appear as $t_i/(1+t_i) = (1 - \bar{b} - \bar{b}\varphi_i) / \bar{\varepsilon}_i = (1 - \bar{b}r_i) / \bar{\varepsilon}_i$, in which $\varphi_i = r_i - 1$ and $\bar{\varepsilon}_i$ is the aggregate demand elasticity of good i , which highlights starkly the equity-efficiency tradeoff. When distributional considerations are not an issue, this formulation further reduces to the partial-equilibrium, inverse elasticity rule.

We will develop yet one more optimal tax problem, that of the optimal choice of taxes on savings and labor income. The consideration of savings brings an intertemporal dimension to the problem, and the income tax raises the issue of labor supply, so a model with consumption in two periods and labor supply is a natural vehicle. We abstract from distributional considerations, however, and use a representative consumer again, whose utility function is $u(c_1, c_2, L)$. The consumer lives for two periods, working in the first and living on savings in the second. The budget constraint then is $c_1 + c_2/[1 + (1 - t_r)r] = (1 - t_w)w(T - L) + A$, where t_r and t_w are tax rates on interest and wage income, T is total time available in the first period, L is leisure, and A is asset income. The consumer maximizes utility (or rather, we as analysts of the problem work through the implications of his maximization by studying

the first-order conditions, in which he varies c_1 , c_2 , and L). Again in our role as analysts, we can use the budget constraint and the three first-order conditions to solve for the demand functions for consumption in the two periods and first-period leisure. Having found those, we substitute them into the utility function to get the consumer's indirect utility function, $v(1 + (1 - t_r)r, (1 - t_w)w, A)$, where the first two terms in the indirect utility function are the relative price of second-period consumption and the wage, both in terms of first-period consumption. Next – again as analysts – we have the tax authority adjust t_w and t_r to maximize the indirect utility function subject to the budget constraint, $t_w w(T - L) + (t_r r c_2) / \{ [1 / (1 + r)](1 + n) \} = \bar{R}$; the term $1 + n$ in the denominator of the revenue from savings accounts for a smaller number of people in the older generation.

Taking first-order conditions, dividing the condition for t_r by that for t_w , and substituting from the Slutsky relationship, we get the following optimal tax rule for the interest and labor income taxes: $t_r r(\varepsilon_{2h} - \varepsilon_{22}) / (1 + n) = t_w(\varepsilon_{hh} - \varepsilon_{2h}) / (1 - t_w) + (r - n) / (1 + n)$, in which supply elasticities of labor are used in place of demand elasticities for leisure (the ε_{2h}). On an efficient growth path, the rate of interest would equal the population growth rate, and the last term uses the tax system to correct for any inefficiency that might arise with the level of the interest rate; if $r = n$, the last term disappears, and the rule can be written as $t_r r / (1 + n) = [(t_w / (1 - t_w)) \cdot (\varepsilon_{hh} - \varepsilon_{2h}) / (\varepsilon_{2h} - \varepsilon_{22})]$, which is strikingly similar to the rule for two commodities and leisure derived above with only excise taxes. A high labor supply elasticity ($\varepsilon_{hh} > 0$) favors lowering the ratio of the labor income tax rate to the tax on interest income; and a low own-price elasticity of demand for second-period consumption would raise the optimal interest-income tax rate relative to the wage-income tax rate. If first-period leisure and second-period consumption are complements ($\varepsilon_{2h} > 0$), the tax authority should levy a relatively higher tax on savings. If we assume the cross-effects to be zero, some further implications of this tax rule appear clearly: $[t_r r / (1 + r)] / [t_w / (1 - t_w)] = -\varepsilon_{hh} / \varepsilon_{22}$. If

labor supply is completely inelastic ($\varepsilon_{22} = 0$), the optimal tax on interest income is zero, and we could tax labor whatever we wanted or needed to raise the target revenue without distorting labor supply.

Optimal tax rules, with redistribution, exist for income taxes, but the potential for lifecycle variation in an individual's relative position in the overall income distribution substantially complicates an already complex set of calculations. Under some conditions of commodity demand and labor supply, an income tax would have better efficiency and distributional properties than would a set of excise taxes or a combination of an income tax and a set of excise taxes. However, those conditions are quite restrictive, and it appears that there are good reasons for maintaining both taxes. Another result we should mention, though it is not for income taxes, is the general undesirability, on efficiency grounds, of excise taxes on intermediate goods, although situations may arise to outweigh the general result, such as inability to tax an important final output.

Optimal tax rules are particularly sensitive to the tax instruments available to the tax authority. They are also quite complicated accounting devices even with simplified specifications of the details of many real-world tax systems for the purposes of modeling. In the past several decades, increases in computational ability have encouraged renewed interest in attempting to apply guidelines, at least broad ones, to tax reforms, particularly in developing countries. It is also obvious that the data requirements for determining the parameters of an optimal tax system are quite demanding, requiring consumer budget surveys and econometric (statistical) estimation of parameters. Some shortcuts are available, and the practitioners' literature is cautiously optimistic about the prospects for this theory's ability to improve an economy's and society's wellbeing through improvements in its tax system. Perhaps they are correct to be optimistic, but without substantial quantitative information, not many qualitative conclusions emerge yet, other than reinforcement of the intuition about the significance of own-price elasticities of demand,

complements to leisure, and the redistributive significance of lowering taxes on commodities that, on strict efficiency grounds, would make excellent revenue collection vehicles.

6.3.5 *Other revenue sources*

We have devoted as much attention to taxes as we have because of their relative importance in governments' array of instruments for raising revenue. Three other categories of revenue warrant brief discussion here. First is debt. Government can borrow, either on open markets, competing with private borrowers, or via strong-arming some wealthy and vulnerable nobles, for example, the liturgies in Classical Greek cities. Many kings have had their private bankers. Debt, however, is not really a permanent substitute for taxation but is, at best, only a revenue smoothing device (this ignores business-cycle stabilization purposes of changes in government debt). The choices government has are to tax now or later. The question, known as Ricardian equivalence, arose recently whether public debt could even affect the timing of net expenditures, or if taxpayers, having their subsequent generations in their utility functions, fully discounted their present wealth by the net present value of the future tax obligation required to repay the debt. It seems as if very special conditions are required for Ricardian equivalence (between taxing now or taxing later) to occur, so debt most probably does change the time profile of revenue.

The second category of revenue source is user fees. Some public goods with imperfect nonexcludability can have their use monitored and users can be charged for their consumption. Roads, port facilities, water supply, libraries, and other public goods along these lines typically are able to charge prices for at least some of their use. However, many of these goods or services are provided at decreasing average cost, which causes a pricing problem we will address further in section 6.5. User fees are examples of pure benefit-principle taxes, which also can give them a regressive character, although as we will see, Ramsey pricing (the second-best pricing framework – see section 6.4 – comparable

to Ramsey taxation principles) can accommodate distributional as well as efficiency goals. As a revenue-raiser, user fees generally are restricted to covering the costs (if they even accomplish that) of the activities they price. They do not leave the government with much discretionary revenue.

The third category of revenue source is government production of private goods. We will deal with this subject in greater detail in section 6.5 as well, but its value as a net revenue source warrants a word here. Public production of private goods requires some explanation as to how government competes with private producers of the same or competing goods – whether by enforceable (or partially enforceable) mandate, exercise of monopoly power after first entry, or through unrestricted competition with private producers. These conditions will affect the amount of markup over marginal cost (“profit”) the government is able to squeeze out of its private-goods production efforts. As with user fees, production of private goods provides only a net revenue, which could offer only a small margin of flexible revenue.

Despite efficiency losses in taxation, and the additional administrative costs of collecting and processing the information necessary to enforce collection of various taxes, taxes are superior revenue generators than the alternatives. Even if marginal efficiency losses (not administrative costs) reach 50%, that is probably a far better profit margin than user fees and production of private goods.

Another source of public finance in Graeco-Roman times was private munificence, or euergetism (Garnsey 1988, 82–84; Migeotte 1997, 38–43; 2009, 149–150; Zuiderhoek 2009; Müller 2011, 331–332). Today we would be tempted to call it philanthropy: you, having more money than you know what to do with, pay for a building that the city fathers can't afford to build and have it named after yourself. In times of food crises, wealthy private citizens might purchase grain and sell it to a city's citizens at something below the crisis-level price, which still may have left them in the black on the transactions. The proportion of public funds coming from such

sources were probably not great and were erratic, sometimes responding to crises, such as the food supply disruptions, sometimes to status competitions among nouveau riche. Zuiderhoek (2009, 332–33, n. 26) notes that in his study of euergetism in Asia Minor during the Roman period, only 3% of the benefactions in his data base went to supplement urban food supplies.

Another category of public funds in Classical Athens was liturgies, a system by which the wealthiest citizens financed and personally supervised sometimes considerable portions of the operations of government (Davies 1967; 1971, xvii–xxxi; Christ 1990; Jones and Rhodes 1996). Athens' magistrates accepted volunteers and then, to fulfill the required liturgies, assigned the remainder to those nonvolunteers they considered most capable of carrying them out. The Athenian navy was financed by trierarchs who paid for a trireme (the advanced warship of the time) and its crew. Otherwise liturgies financed theatrical performances and festivals.

6.4 The Theory of Second Best

The theory of second best deals with optimal policies, in general, in the presence of irremediable distortions in some markets (Lipsey and Lancaster 1956–1957). We have hinted at the issue earlier: once one of the efficiency conditions for a Pareto optimum is ruled out (it cannot be attained for one reason or another), it is not necessarily a good thing to maintain all the other marginal conditions. We offer a brief, formal development of the theory here.

The usual optimization problem (that is, one in which a Pareto optimum is possible) has us maximize, say, a utility function, $u(x_1, \dots, x_n)$ subject to a transformation relationship, $F(x_1, \dots, x_n) = 0$, that defines the frontier of production possibilities (sometimes this is just the budget constraint). The first-best solution is to maximize u subject to F , from which the first-order conditions are $\Delta u/\Delta x_i = \lambda(\Delta F/\Delta x_i)$ for all the x_i and $F(x_1, \dots, x_n) = 0$. The first of these conditions is the requirement for efficiency that the rate of substitution equal the rate of

transformation: $MRS_{ij} = (\Delta u/\Delta x_j)/(\Delta u/\Delta x_i) = (\Delta F/\Delta x_j)/(\Delta F/\Delta x_i) = MRT_{ij}$.

In circumstances that characterize second best, there is another constraint. Suppose that for one of the goods $(\Delta u/\Delta x_k) = k(\Delta F/\Delta x_k)$, where $k \neq \lambda$. There may be a politically imposed pricing constraint on good k . Now the maximization problem is $\mathcal{L} = u(x_1, \dots, x_n) - \lambda F(x_1, \dots, x_n) - \mu[\Delta u/\Delta x_k - k(\Delta F/\Delta x_k)]$, and the first-order conditions are $\Delta u/\Delta x_i = \lambda(\Delta F/\Delta x_i) + \mu[\Delta(\Delta u/\Delta x_k)/\Delta x_i - k(\Delta(\Delta F/\Delta x_k)/\Delta x_i)]$. The marginal rate of substitution between any pair of goods x_i and x_j is accordingly $(\Delta u/\Delta x_i)/(\Delta u/\Delta x_j) = \{\lambda(\Delta F/\Delta x_i) + \mu[\Delta(\Delta u/\Delta x_k)/\Delta x_i - k(\Delta(\Delta F/\Delta x_k)/\Delta x_i)]\} / \{\lambda(\Delta F/\Delta x_j) + \mu[\Delta(\Delta u/\Delta x_k)/\Delta x_j - k(\Delta(\Delta F/\Delta x_k)/\Delta x_j)]\}$. The relative prices of goods x_i and x_j should equal this lengthy expression on the right-hand side of the last formula. There is no reason the marginal rate of transformation should equal the same value. Once the efficiency conditions somewhere in the economy are no longer met (in this case, for good x_k), it is no longer optimal for relative prices of other goods (not involving x_j) to equal their relative marginal costs.

When you think about it for a minute, this is a potentially catastrophic finding for a science developed to offer policy guidance on how to improve wellbeing. Making recommendations about how to modify tax rates or an entire tax system (tax reform) is a case in point. Part of the problem is again computational, part again informational. However, in the 40 years since this result was first published, the subject, and how to get around the problem with various approximations, has been the subject of intense research, and much has been learned. Sometimes groups of goods can be partitioned off “safely” (that is, without serious distortion to their relative proportions in consumption or to relative proportions of goods outside the group), and marginal cost pricing can be applied to a group of goods with anticipation of making a welfare improvement, even if such pricing does not yield the maximum welfare level that could be attained with better information. These recommendations sometimes are called “third-best” policies or actions. Nonetheless, the legacy of second-best

theory is to be careful when removing distortions in the economy on a piecemeal basis: you could make things worse off than they were before if you aren't careful.

6.5 Government Productive Activities

We want to emphasize the distinction between public provision of goods and public production of goods. Public provision refers to the government's seeing that public goods are supplied to the population. Government can contract with private producers to produce weaponry for national defense, for public construction projects such as roads, harbor facilities, public buildings, city walls, irrigation facilities, etc. Alternatively, the government could literally own the production capabilities for these items and make all the production planning decisions itself (planning how to produce what is already decided upon, as contrasted with planning what to produce and how much).

The goods we have offered as examples for government provision are all public goods. Governments sometimes provide goods with private characteristics, items ranging from foodstuffs to, in contemporary societies, electricity. The provision of food may be in fact a particular type of public good – the merit good, something of which society wants everyone to consume at least a minimum amount; depending on private sources to supply this public good may be leaving the matter to a greater degree of chance than society would want to take. The Classical Greek city states are well known for this concern. Such private goods may be produced by private producers either under contract to government or selling to government agencies as to any private demander. Electricity (an ancient good with comparable public characteristics might have been water) is certainly an excludable good: the electricity one of us consumes is electricity that the rest of us cannot consume. However, electricity can be produced efficiently on a very large scale, giving that industry a declining average cost curve which, as we have pointed out

several times, can cause efficiency problems with private supply.

Finally, government can *produce* – as contrasted to simply directing their redistribution – purely private goods. That endeavor, in fact, can be said to have been the great social experiment of this century, but if Sumerologists such as Anton Deimel and Anna Schneider are correct (and possibly even if they're not), it may be a repeat experiment. At any rate, the subject of public production of pure private goods, as well as public provision or production of public goods, is of clear interest to students of the ancient societies of the Mediterranean and the Aegean.

This section begins with the economics of public production and pricing of excludable goods. The second subsection deals with the supply of public goods, including social mechanisms societies may use to decide what they think on those matters. The final subsection deals with public investments and criteria for evaluating them.

6.5.1 Public production and pricing

Since pricing is critical to efficient production, we begin this subsection with an analysis of public pricing policies. The different budgetary conditions of genuinely private producers and public firms or agencies form the second topic. The economics of the public firm leads naturally into a discussion of public economic planning.

Pricing and production in declining-cost industries

Publicly produced private (excludable) goods in the noncentrally planned societies of the contemporary world tend to be those with large fixed costs and declining average cost curves: electricity, water supply, jet aircraft. These goods are not uniformly produced by government around the world, but where they are not, government frequently regulates some aspects of their activities or otherwise has a major influence in the business, such as through purchases, subsidies, or both. These activities typically have declining average costs in the region of production, which means that the marginal-cost pricing required for efficiency will not cover average cost. In such

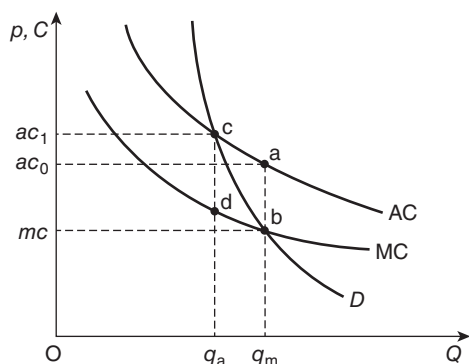


Figure 6.15 Marginal cost pricing in a declining-cost industry.

industries, private producers who tried to price at marginal cost would go out of business. Those who didn't would restrict output below the socially efficient quantity and earn potentially sizeable rents ("profits") which would attract the attention of government anyway.

Figure 6.15 shows this predicament. Average and marginal cost curves are AC and MC. The demand curve, D, cuts both average and marginal cost curves. Marginal cost pricing would supply quantity q_m of the good but would lose $ac_0 - mc$ on each unit produced, for a total loss of rectangle $(ac_0)ab(mc)$. The largest output this firm could sell at a profit is q_a , determined by the intersection of the demand and average cost curves. Consumers would be willing to pay an additional amount equivalent to triangle cdb , on which units the producer can cover the variable costs but not the fixed costs. This is the central problem of public sector production and pricing. One solution is for the public firm to price according to marginal cost and for the treasury to subsidize the fixed costs with revenue from lump-sum taxes. This is neat on paper but hasn't worked out quite so straightforwardly in practice.

Ramsey pricing

The same apparatus from the Ramsey tax model can be used to construct second-best prices for declining-cost industries that will let them price as close to marginal cost as they can while

still applying a mark-up over cost sufficient to cover average cost. There are many alternative formulations of the Ramsey pricing model. The first choice we have is what objective function to maximize. Taking a restrictive view of the problem, we could maximize the sum of consumer surplus of the firm's customers and the producer surplus of the firm. That can give us a second-best, efficient solution, but it does not address any distributional concerns we might have about various consumers' ability to purchase the product. Alternatively, we could have the firm maximize the indirect utility function of a representative consumer, which still does not address distributional concerns. Another alternative would be to maximize a social welfare function that would let us, for all practical purposes, weight the consumer surplus of different individuals differently and adjust the price of a single-product firm's output or price an array of products of a multi-product firm according to some combination of efficiency and distributional desiderata. The basic constraint is that the firm cover its costs from either its own revenues or a combination of its revenues plus a subsidy. Other constraints could be placed on the firm as well.

For simplicity, we show first the maximization of the representative consumer's indirect utility function, when the firm must cover its costs with its own sales. The indirect utility function is $v(\mathbf{q}, p)$, where $\mathbf{q}$ is an array of prices of private goods and p is the price of this public firm's product. The firm's budget constraint is $pQ - c(Q) \geq \Pi$, in which Q is the output of the public firm, $c(Q)$ is its cost function, and Π is a target level of profit, which may be positive, zero, or negative. If we wanted to offer the firm a subsidy as well, we would add it to the left-hand side of the constraint. We do not show the Lagrangean, but the first-order condition from adjusting the public price, p , optimally is $(p - mc)/p = [1 - (\alpha/\lambda)](1/\epsilon)$, in which α is the marginal utility of income, λ is the marginal cost of public funds (the Lagrange multiplier on the firm's budget constraint), and ϵ is the elasticity of demand for the publicly produced good. The markup will be in

inverse proportion to the demand elasticity of the product. With a multiple-product firm, the Ramsey rule for the ratios of different prices would be $s_i(p_i - mc_i)/p_i = s_j(p_j - mc_j)/p_j = \dots s_n(p_n - mc_n)/p_n = \lambda$, where the s_i terms are coefficients containing the cross-elasticity information. To give a flavor of the content of these coefficients, in a two-product case, $s_1 = (\epsilon_{11}\epsilon_{22} - \epsilon_{12}\epsilon_{21})/(\epsilon_{22} - \omega\epsilon_{21})$, in which ω is the ratio of sales of product 2 to sales of product 1 —: p_2Q_2/p_1Q_1 . When the cross-elasticities are zero, the s_i terms simplify to the own-price elasticities ϵ_{ii} , yielding the inverse elasticity rule. In either case, the percentage markup across products is not uniform, but depends on the sensitivity of each product's demand to prices.

Using a social welfare function that permitted us to modify efficient prices with weights to accomplish distributional goals, the Ramsey rule would, for a single product, be $(p - mc)/p = [1 - \bar{b}(1 + \phi)]/\epsilon$, in which $\bar{b}$ and ϕ have the same definitions as in the discussion of optimal taxation: average social marginal valuation of income and the covariance between social marginal valuation of income and consumption of the publicly produced good. It can be seen from this distributional Ramsey rule that it is possible for distributional concerns to make an optimal public enterprise price less than marginal cost.

By now, you probably will have noticed a connection between government and the "private sector" in the public pricing problem that is not present in optimal tax systems. If there are demand interdependencies (cross-elasticities) between the publicly produced goods and privately produced private goods, there is a linkage between the public firm's pricing policy and private producers' profits. This may provide government an indirect way of taxing profits, but it also may provide political grounds for additional constraints on the Ramsey pricing problem.

The Ramsey pricing rule permits the public firm to discriminate among the consumers of different products according to the demand conditions for each product, but it does not incorporate the ability to discriminate among different individuals consuming the same product. Why might it be desirable to charge different individuals different

prices for the same product? Before addressing that question, it would not be productive to charge different prices for re-salable goods; individuals qualifying for the lower price could re-sell the good at the higher price, re-purchase at the lower price, and so on, till the system broke the firm financially. The two principal reasons for charging different customers different prices are different tastes (and hence demands) among the customers and different costs of serving them. If the different customers can be persuaded to pick their price, plus whatever tie-ins are necessary accompaniments of the different prices, each customer will be satisfied with his or her price.

Since we have introduced the subject of income redistribution through pricing of publicly produced goods, it is a reasonable opportunity to show the standard analysis of the inefficiency of redistribution in kind. Figure 6.16 shows the welfare loss associated with in-kind transfers relative to fungible (cash) transfers. Indifference curve I_0 shows a consumer's equilibrium choice of combinations of goods x_1^0 and x_2^0 , given his restriction to budget AB and the relative prices embodied in the slope of AB . Now, we give this person $x_2^1 - x_2^0$ of good x_2 . At the same relative prices in AB , this puts him at point a on a budget line $A'B'$. However, at that budget level, this consumer would prefer point b , where his indifference system has a tangency with the available marginal rate of

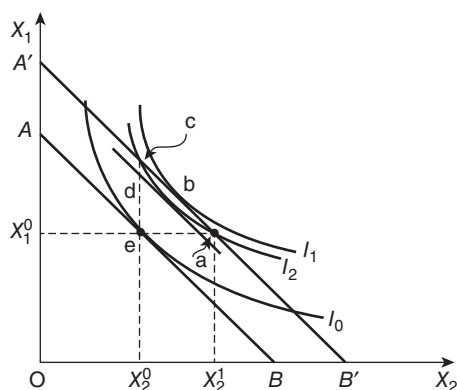


Figure 6.16 Inefficiencies of redistribution of income in kind.

transformation (the slope of the budget line). His indifference system cuts the budget line at point a, where the transfer in-kind placed him, which means that he could do better moving toward the center of the two intersections of his indifference curve I_2 with $A'B'$. The vertical distance between point b and indifference curve I_2 is a measure, in terms of good x_1 of welfare cost of this transfer: whereas the transfer we gave him cost us the equivalent of distance ce in terms of good x_1 , in the recipient's valuation, it was worth only distance de in terms of x_1 , the welfare loss being distance cd. The moral of the story is that there will always be a welfare cost of redistributions in kind rather than in some substance that the recipient can convert into the consumption bundle that will give him the greatest satisfaction.

Readers, fairly, might wonder how prices such as Ramsey prices, that use so much information on preferences and technologies of diverse agents, not to mention the theoretical concepts that tell how to use the information, could be developed, even in approximate form, in even the most sophisticated of the ancient Mediterranean and Aegean societies. (They will probably have the same question about multiple-part pricing, just below, so we'll address the issue here.) I'll take the long-winded route in making my suggestions. First, the proximity in time and location of production and the experience of marginal cost lends some intuitive character to marginal-cost pricing even in the absence of markets: "How much d'ya want for this?" (Aside to self: "Man, am I tired. That last one wore me out.") "Two bags." "Hmpf. Seems a little pricey to me. Well, gimme four of 'em." Take the same type of proximity to a larger production organization. Second, the concept of the demand elasticity has considerable intuitive appeal. The notion that some goods are things that people are going to consume a lot of, regardless of its price, and the observation that these goods tend to be things like food and water, followed by the simplest of clothing and shelter, surely has been with us for millennia. Third, concepts about the proper treatment of the poor, combined with political savvy about where most of your soldiers are going to come from,

also are not new. Both ethics and political commonsense would restrain – but not necessarily prevent altogether – the raw exploitation of low demand elasticities for necessities. Fourth, the concept of unequal influence among individuals, with influence tending to be weighted by wealth, because of both wealth's immediate purchasing power and its signal regarding the capabilities of the individuals possessing it, again, surely cannot have been missed by kings who scrambled their way to power in the ruthless times of the ancient world – nor by the elective politicians of the later, Classical Greek world. So, neither would it be desirable (smart) to redistribute income so heavily as to irremediably offend the people who can hurt you much sooner than the mass of common soldiers. So we have some reasonable grounds for suspecting that the full array of social forces required to make the adjustments to marginal-cost prices that Ramsey prices entail may well have existed in most societies for a long time. The Ramsey price model, derived for a social welfare function, is a contemporary codification of the principles underlying behavior that has characterized societies for quite a while. Its codification lets the calculations be made more easily and gives scholars and public officials charged with setting public prices clearer focal points to debate. The principles the model codifies are not new; the clarity of the understanding of them may be. And for the reader who would be somewhat more satisfied with a more formal account of how such complicated prices might be put into practice, we can mention that a Ramsey price structure can emerge from negotiations among coalitions known in theoretical terms as cooperative games, the negotiations being the mechanism that yields the pricing structure. The prices reflect the relative bargaining strengths of all subcoalitions of consumers of the firm's products (Spulber 1989, 7–8. 179–247).

Multiple-part pricing

Multiple-part, or nonuniform, pricing in general is a combination of some fixed charge for consuming any at all of the good plus a price per unit of consumption. A two-part price will have only a fixed charge and a variable charge per unit

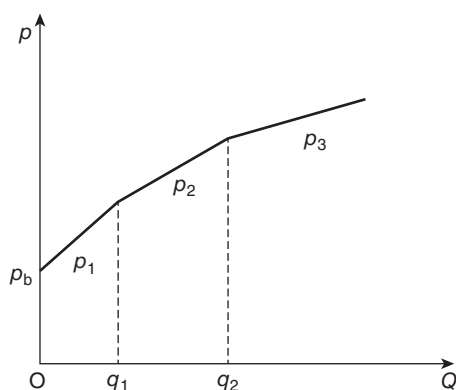


Figure 6.17 Multiple-part pricing.

of consumption. Multiple-part prices will offer different rates over different ranges of quantities purchased. Frequently, larger purchases will get lower unit prices. Figure 6.17 shows a four-part price schedule; the first part is the fixed charge of p_b . Consumers purchasing from zero to q_1 units will pay price p_1 per unit. Purchases greater than q_1 but less than q_2 will pay p_b plus p_2 per unit, and quantities greater than q_2 receive price p_3 per unit, plus, of course, the fixed charge, p_b . The price schedule can be tailored to let the fixed component vary among consumers as well as the variable component. Such a structure is particularly useful for declining-cost activities. Multi-part pricing also can help firms adapt to random variations in demand by shifting some of the cost of risk onto the consumers who contribute to it.

Nonuniform pricing provides more flexibility to match marginal production cost to marginal consumer benefit. Such a system allows consumers to identify themselves according to demand. Compared to a uniform Ramsey price that exceeds marginal cost, the multipart price schedule can make both consumers and the firm better off.

The structure of public enterprises

We will not present any specific models of public enterprises or government agencies that produce goods for sale. Instead we will point out some important parameters that may vary across cases

but which are influential in determining the behavior of the firm. First is the character of the firm's budget: is it fixed, soft (in the sense that it is negotiable upward), or unlimited? Is its budget influenced by various indicators of its performance? Are those indicators unbiased measures of the firm's efficiency or the consumer plus producer surplus it generates? Second, does it need a budget? Can it keep any profits it makes or must it turn all profits over to the treasury? Third, in the absence of the ability to retain profits, what incentives do the firm's managers have? Do they have the flexibility to operate the firm in such a way as to maximize their own (constrained, of course) utility at some cost to the potential welfare of the government? If pay raises and promotions are the official incentives of the managers, how do the values of those payments compare with unofficial benefits that would not maximize pay raises and promotions? Third, what is the structure of oversight of the firm? How accurately and cheaply can its efficiency be monitored by disinterested outsiders?

In other words, "What's in it for the manager of a public firm or agency?" From the distance of four or five millennia, it is easy to suppose that it was just a matter of "off with his head" for perceived malfeasance or even for less-than-maximum efficiency, but those are costly and not necessarily efficient incentive systems. The alternative assumption of extreme altruism on the part of ancient managers is no more satisfactory. The absence of public statements to the effect that the leader of an official expedition to Punt traded extensively on his own account as well as collecting specimens for Hatshepsut is not evidence that he did not (Breasted 1906, 102–122 §§ 246–295). Why advertise?

Public economic planning

There is always planning of one sort or another. If it is not explicit, it will be implicit. Someone running a firm today would find difficulty in not thinking at all about what it will take to keep operations running tomorrow. Planning can be highly centralized (say, the king does it all and all other administrators just carry out the plans) or quite decentralized, with general directions

given to firms and agencies, but implementation details left to firm managers and agency administrators. The same amount of information may be required to plan for future operations from either a centralized or decentralized perspective, but the mechanisms for assembling the information in the locations where it is most valuable will differ, as will the costs.

There is a relationship between information, the completeness of markets, and the tasks for a planner. Consider the concept of the time-dated, contingent commodity, such that bread today, bread tomorrow, and bread the day after tomorrow are different commodities and hence comprise different, if closely connected markets. Each day's bread may depend on the weather (one market for rain, another for shine), activity in other markets, or other "states of the world" that cannot be predicted perfectly. With a sufficiently large number of possible states of the world (more plausible, or at least certainly more interesting, for commodities like wheat and barley than for bread!), any particular one of them is highly unlikely to occur, and it will not pay to establish a market for it. If the array of markets were complete, it would be possible to make all decisions for the future at some initial time – say, on the first day of a new king's reign, or maybe after the end of the first week or month, giving him time to plan. With incomplete markets, "recontracting" will be desirable at any number of future dates; that is, the planner will want to change his mind as the states of the world unfold differently from anticipations about them.

Now, there is an inherent tension between the existence of the planner and the existence or reliance upon markets. In the 1930s, Oskar Lange proved that an omniscient planner would simply replicate the resource allocation that perfect markets would yield (Lange 1936–1937a, 1936–1937b),²⁴ but that result depends on (i) the coincidence of the planner's preferences and the preferences that might be aggregated from the population and (ii) the costs of assembling and processing the information that markets assemble and process, but doing so outside, or without the aid of, markets. A very large amount of information must be assembled *and coordinated* before

even a feasible allocation of resources can be determined, much less an optimal or efficient one. What do we mean by a "feasible allocation?" Try an example. Suppose we planned on making five thousand spears and two hundred chariot wheels. In our measure of labor inputs, a spear requires two units of labor (ignore substitution possibilities for the example) and a chariot wheel requires ten. We have ten thousand units of labor available. We have an immediate problem of infeasibility. Surely we could construct more intricate and interesting examples of infeasibility, involving less transparent tradeoffs that the planning goals do not accommodate, and revealed by physical shortages of inputs, queues for outputs in physical shortage, the occasional physical surplus of stuff that nobody wants or can use before it rots or dries up, etc. But such are the embodiments of the concept of the infeasible allocation.

The planner has the requirement to assemble information on technologies, production requirements, and total factor supplies (knowledge of tastes is unnecessary since the central planner may substitute his own preferences) – plus contingency plans for the weather and other acts of nature and people outside his or her own "control." To get a rough idea of just how much information this might be, consider a rather abstract central planning problem with an example taken from Starrett (1988, 26–31). Our planner (the monarch?) sets desired (by him or her) consumption levels C_i of goods 1 through n and assigns production choices y^j involving technologies superscripted " j ." His problem is to choose production levels that satisfy the following relationship: $C_i \leq \sum_j y_i^j$ for all commodities i . One way of doing this in practice is to leave consumption of one of the commodities as a residual after satisfying all the other target levels. Let's call this particular commodity C_0 . Then the maximization problem is: Maximize C_0 subject to $C_i \leq \sum_j x_j a^j$, $x_j \geq 0$ being the scale at which the activity is operated and a^j is an activity vector characterizing how inputs are combined, for each commodity i , where subscript i is numbered from 1 through however many consumption targets we have (we might have separate consumption targets for different groups, and even for some

specific individuals). The x_j and a^j terms break the information contained in y^j into separate components representing the level at which the production activity must be conducted and the array of separate inputs required to support that level; that is, we could disaggregate variable a^j further into labor, animal traction, seed, water, land, and tools for the production of field crops.

Let's look at the information requirements for solving this problem. There is one constraint for each targeted commodity, and each activity or input in the a^j collection is a separate variable. And this is for only one time period. The information requirements rise linearly as a multiple factor of the number of goods whose consumption levels are targeted. Then there is the matter of coordination between producers of goods that are used as inputs in other activities. This amounts to possibly several other pieces of information (multiplied by two for origin and destination) for each such intermediate good. This knowledge is dispersed among economic agents. Given efficiency incentives, these agents transmit this knowledge among themselves and process it with computations that result in allocative decisions. Because the knowledge is dispersed, agents are able to get away with transmitting false information when doing that will benefit them. These informational asymmetry problems are difficult to handle in markets that let individuals try to determine their own benefit levels. In systems of planned consumption, the incentives to transmit false information will arise as protective mechanisms and are just one more information problem for the planner to overcome.

If the commodities are bundled in some fashion in order to lower the information costs for the central planner, the information requirements do not go away but get shoved down to some lower-level administrator, who still has interagency coordination problems, leading to a hierarchical structure, with information requirements of its own. If we are prepared to believe that governments directed large portions of their "countries'" economies, we must accept either that they developed solutions to these information problems or that they lived with the consequences, which could have ranged from endemic

physical shortages and wasted surpluses,²⁵ to black markets to political unrest. These are rationing systems, and the main problem with rationing systems is that individuals frequently try to circumvent the official quantity rules and trade in black markets. In this sense, the ability for markets to arise alongside rationing systems is just as much a constraint on a planning system as a transformation constraint (budget or production function) is on an individual maximization problem. One of the largest benefits of markets as allocation devices is that they're not as bad as the next best mechanism – incompetent and corrupt politicians and administrators, a point well expressed by Hammond (1990, 13–15).

6.5.2 *The supply of public goods and social choice mechanisms*

Recall our discussion of the demand for public, or you might want to call them collective, goods in section 6.2. Figures 6.1 and 6.2 showed the vertical aggregation of the individual marginal rates of substitution (or demand, or marginal benefit, curves) for an inexhaustible, nonexcludable public good. Recall also the individualized tax payments $t_i G$ that would give us just the required revenue to pay for the amount of the public good that everyone wanted, which we called Lindahl prices and said were very hard to collect because everybody had an incentive to understate his preferences once he learned the tax game. This is the famous free-rider problem. Nothing we have discussed between then and now has gone any way toward resolving this problem.

Let's step back a moment and restate just what this problem is. We're confident that there really is some amount of public goods that a collectivity of citizens (subjects) wants. We're also confident that we can show analytically what that amount is and prove that that amount will make everybody as well off as they can be made without making anybody else worse off. One of the problems in public good supply is that lots of people don't really mind making someone else worse off (or even better, lots of people they don't know just a tiny bit worse off, so they'll hardly notice, if at all) in order to make themselves better off. It's difficult

to do that with private goods, but the combination of inexhaustibility (not even pure inexhaustibility) and nonexcludability make it possible in public goods – in fact, difficult to avoid.

We can go in two directions from these few facts. One is positive and would examine how the quantity determinations of public good supplies have been made in practice, both recently and long ago and in quite different political conditions. The other is normative and investigates how “good” public good supply decisions might be made, where “good” entails a combination of efficiency and ethical norms. We need not try to keep the two directions apart, since what we learn in one can inform the other. Studies of how these supply decisions have been made in the past can give us insights into the mechanics involved, and study of the relationships between allocative efficiency of various procedures and their ethical implications offers us evaluative insight on the tradeoffs that previous and contemporary societies have made and continue to make. We begin with the normative direction, which actually contains quite a bit of positive analysis of how various mechanisms work. Before pushing off into the open waters, we point out that, whereas we have discussed this issue in terms of how much of a public good will actually be supplied, the current subject is really the revelation of the demand for public goods. It is a matter of the “quantity supplied” of public goods inasmuch as the supply curve is fixed, and we are now trying to learn what we can about how the quantity actually purchased is chosen.

The optimal provision of public goods financed by taxes

Before proceeding with how a society can go about deciding how to decide upon the allocation of public goods, let's take one last look at the amount of a public good that would be optimal for the government to provide – that is, the answer that they “ought” to come up with. In section 6.2, we showed that a public good should be provided in a quantity that would make the sum of the demands for it (marginal rates of substitution between the public good and some numeraire good) equal the price of the good.

That was before we introduced the distortionary effects of taxation that raises the revenue. Clearly, the Lindahl public good provision rule should account for this effect but, contrary to simple intuition on the matter, it is not necessarily the case that we should simply elevate the direct cost (possibly the price the government has to pay “on the market” to hire contractors to provide it) by the marginal cost of public funds. That would reduce the provision below what the unadjusted Lindahl rule indicates.

Let's start with the representative consumer, whose utility is a function of a composite private good, Q , the inexhaustible and nonexcludable public good, G , and leisure. The individual budget constraints are $(p + t)Q = L$. The consumers' aggregate budget constraint is $p \sum_h Q^h + p_G G = \sum_h L^h$, where L^h is labor (the tax doesn't go in the aggregate budget constraint because the sum of the tax levies is equivalent to p_G). Since the representative individual assumption implies identical consumers of the public good, the Lagrangean the government will maximize is $\mathcal{L} = H \cdot U(Q, G, T - L) - \lambda(pHQ + p_G G - HL)$. The first-order condition is $H(\Delta U / \Delta G) / (\Delta U / \Delta Y) = [\lambda / (\Delta U / \Delta Y)] \cdot [p_g - tH \cdot (\Delta Q / \Delta G)]$, where $\Delta U / \Delta Y$ is the marginal utility of income. The last term on the right-hand side of the expression says that if the cross-effect between the public good and the private good induces consumers to consume more of other taxed goods, the tax that needs to be raised to finance this public good is less than it would be in the unadjusted Lindahl equilibrium. However, the government still needs to behave as if the price for which it can furnish the public good, p_G , were really λ times higher than it looks on the invoice, to account for the marginal cost (deadweight loss) of taxation. If the cross-effect between the public good and other taxed private goods is zero, we are back to the simpler Lindahl rule. If we had maximized a social welfare function we could have obtained welfare weights such as we obtained in the analysis of optimal taxation, and the Lindahl rule would be adjusted to $\sum_h (\beta^h / \bar{\beta}) \text{MRS}^h = p_G$, or $\sum_h \text{MRS}^h (1 + \varphi_G) = p_G$, where φ_G is the covariance between the relative marginal social valuation of income of each individual h and his

relative marginal rate of substitution between the public good and a numeraire – again one of the terms we encountered in optimal taxation to address distributional issues.

Voting on public goods: the Arrow impossibility theorem

The single, most important focal point in the literature on social choice mechanisms is Kenneth Arrow's impossibility theorem, which is an analysis of the potential for various types of voting to help a society reach a Pareto optimal allocation of resources in collective goods (Arrow 1963). Arrow's theorem says, essentially, that we must choose between economic efficiency and what we today would call democratic political values. We can't have perfect efficiency and perfect democracy. There are many in-between combinations, and much of the literature following Arrow's original publication in 1951 has been devoted to exploring what could be called the second-best options. Arrow asked the question, "Does there exist a collective choice process that (i) can find and enforce efficient allocations of resources, especially when markets can't reach such an allocation (the Pareto-optimality requirement); (ii) will work for all possible combinations of individual preferences (the "unrestricted domain" requirement); (iii) can rank all the alternatives (the rationality, or transitivity, requirement); (iv) rank them efficiently (the "independence of irrelevant alternatives" requirement); and (v) be nondictatorial (the nondictatorship requirement)?" Both the short and long answers are, "No," but the reasons why not are themselves particularly interesting. But, why study social choice mechanisms anyway, if your interest is really in ancient societies? All societies make collective choices, so they implicitly use some social decision function. Whether they were particularly aware of it or not is one matter; our attempts to understand what they did, how, and why, are another matter. Clearer understanding on the parts of contemporary scholars of the various tradeoffs involved in any particular method of reaching collective decisions (or decisions for collectivities) may yield new insights to long-standing questions, raise new questions, and

generally focus the intellectual debate on what is already known.

We consider the consequences of relaxing each of the five requirements of Arrow's social choice mechanism. Relaxing the Pareto optimality requirement either results in a social choice mechanism that is always indifferent between alternatives regardless of individual rankings or it produces a negative dictatorship, a situation in which the mechanism always chooses the reverse of what one agent prefers. If we relax the unrestricted domain requirement, someone's voice is excluded from the collective decision. Trying to compromise on this requirement with majority rule on pair-wise voting ("do you prefer A to B?") cannot be guaranteed to yield a unique majority outcome if more than one issue is considered – for example, the size of the budget; considering size of budget and taxes together could fail to find a stable outcome.

The rationality requirement came under the most pressure from research because it seemed the least defensible of the five elements. Two types of transitivity are at the heart of the rationality requirement, preference transitivity (called P-transitivity) and indifference transitivity (called, predictably, I-transitivity). P-transitivity says that if someone prefers x to y and y to z , he'll prefer x to z . I-transitivity says that if someone is indifferent between x and y , and between y and z , she'll be also be indifferent between x and z . Relaxing these requirements would let a collective choice mechanism simply look for the "best" choice rather than saddling it with the responsibility for completely ordering all the choices. The original purpose of the rationality axiom was to give path-independence, or agenda-proofness, to decisions – that is, to prevent the order in which a vote appears on the agenda of votes from influencing the outcome of the vote. Thus, it blocked the power of agenda-setting. If we relax the requirement of I-transitivity, which seems innocuous enough, we create oligarchies: groups whose agreement with one another over a pair of alternatives is sufficient for the society to choose the group's preferred alternative regardless of other people's preferences. But if this group of oligarchs is not unanimous, any one of them can

veto any alternative, effectively ruling out any collective decision at all. Weakening of P-transitivity, not just eliminating it altogether, can yield some unique decisions over restricted choices, but suffers from not being single-person-veto-proof when the number of pairwise comparisons approaches the number of voters. Altogether, relaxing the rationality requirements does not guarantee a successful social choice mechanism.

Independence of irrelevant alternatives rules out the intensity of preference as a voting criterion. That may seem unreasonable if people have very strong first choices followed at a great distance by their second choices. Weighted, or point, voting is just such an intensity voting mechanism. In the so-called Borda count, with x alternatives and n voters, each voter assigns x points to his most preferred choice, $x - 1$ to the second choice, and so on. The Borda score, which identifies the winning choice with the highest score, is the sum of the points from each voter. The Borda count satisfies the other four requirements, but has its own unattractive feature: revising the scores on just two of the alternatives that were far back in the running (you have to change at least two scores, because the sum of each voter's scores remains the same), can reverse the order of the most and second-most preferred choices. Alternatively, if we dropped the worst option from the running, that also could reverse the scores of the top two choices. This type of count voting is open to strategic voting, in which people do not vote their true preferences in hopes of being able to influence the subsequent agenda.

Finally, we could relax the nondictatorship requirement. However, nothing in the other four requirements guarantees truthful preference revelation to a dictator. Even if a dictator decided to take a poll, he could offer an optimal allocation of public resources on the basis of false preferences, and it's not clear what that buys anybody.

One interim solution that has been used in models of endogenous²⁶ public goods supply is the median voter model. The idea is that in a majority vote, it really takes only one voter to determine the outcome of the vote – the one whose preferences, on a continuum of preferences regarding the issue under vote, are right

in the middle. More formally, the model lets the community demand for a public good be the median of the individual demands for the good. This operates on the demand side of public goods supply, not on the supply side – the vote is on the quantity of a public good to be purchased by a community. The operation of this model requires some strong assumptions, such as no agenda-setting and particularly simple public goods.

Preference-revelation mechanisms

Considerable research has been conducted on what have been called preference-revelation mechanisms – procedures, or sets of rules, that will induce agents to reveal their true preferences. These mechanisms have not really extended into collective choice along the lines of voting, but have offered some benefits to a variety of asymmetric information situations, some important classes of which involve interactions with government in, for example, bidding for leases and for procurement contracts. In winner-take-all bids at the highest price offered, bidders will try to strategize against one another, with the winner frequently finding that he bid more (or less, depending on the contract) than he finds profitable. In general, these mechanisms let bidders share consumer's surplus in ways that prevent bidders from winning things they don't really want in return for telling the truth about their own preferences (or technological characteristics, in the case of procurement bids and other production-related bids). Specifically, they pay the net increase in the sum of consumer and producer surpluses of the other bidders in the market that were identified by the revelation of the supply or demand curves under scrutiny in the auction.

The auction we just described is called a first-price, or English, auction. The highest (or lowest, if the bid is to offer services) bid wins and the bidder pays that price. A variant on this auction is the Dutch auction, in which the auctioneer starts at the highest price, and the first bidder wins, paying that price. Both of these auction structures encourage strategic bidding and falling afoul of "the winner's curse." The so-called

second-price, or Vickrey,²⁷ auction awards the prize to the highest bidder but makes her pay only an amount equal to the second-highest bid. The winner thereby gets the consumer surplus on the deal and has no incentive to either understate or overstate her preferences. A more general class of such mechanisms (they're actually called just "mechanisms") is the Groves–Ledyard–Clarke mechanism, in which individuals are given the incentive to reveal their valuations truthfully by being rewarded in proportion to the sum of all agents' declared valuations.

The structure of such mechanisms is just what would prove useful in solving the free-rider problem of preference revelation in public goods supply, except for two problems. First is the large-numbers problem. With large numbers of demanders, these mechanisms would be unwieldy. Second, they do not guarantee to cover costs, so at the very best, the government still would have to subsidize a number of public goods. To date, their use has been restricted to auctions with small numbers of bidders in both public and private settings.

The positive economics of politics

We promised to start with the normative strand of work in public goods supply, and while much of the work we just discussed is indeed concerned with social optima, much of it also is within the realm of positive economics. The mechanisms work, when its question, "How can we make such-and-such happen?" straddles the positive and normative at the least. In the following paragraphs we will treat issues that may fall somewhat more into the positive part of the house than the normative, but we should emphasize the mutually beneficial interaction among these topics. They are all at the borderlands of economics and politics.

While the work on social choice mechanisms sparked by Arrow's theorem focuses on the viability of different voting mechanisms, and by implication, points to some possible determinants of, or influences on, political structures and interaction patterns that exist in the empirical world, I turn to models that have been motivated more immediately by the desire to understand

current or historical practices. Frequently, apparently wasteful, even corrupt, practices appear to be something close to second-best responses to insuperable constraints, of political power or technology or both.

In the contemporary, Western industrial democracies, much of public spending has been determined in the legislative process. Legislatures make their decisions by voting, but considerable "homework" generally is done prior to a vote even being allowed by the people who control the agendas – the issues treated so abstractly in the social choice literature. An Americanism that apparently has spread at least as far as England is the term "logrolling."²⁸ Logrolling is the common practice in legislative politics whereby legislators enlist support for spending programs they favor themselves by offering to vote for other legislators' spending programs – vote trading, in short. The theory of logrolling is still relatively little developed, compared to other issues in public choice, but some results are worth noting. First, the system may be rife with incentives to cheat, with a game-theoretic structure like the Prisoner's Dilemma, in which both parties would be better off if they didn't cheat, but the incentives to not cheat are symmetric, fostering distrust. However, the repeated-game might be able to solve that problem, and may in fact be a stabilizing contribution of political parties in systems in which individual legislators enter and exit with comparative frequency. The omnibus legislative package is another device to reduce cheating, and it is, in fact, the omnibus spending bill that is the classic logroll. Nonetheless, there seems to be a tendency for logrolling equilibria, when they manage to exist, to over-provide public goods.

The politico-economic constraints on dictatorial choices have not attracted considerable attention in economics. Nonetheless, some mechanisms seem plausible. First, forces of economic scarcity (otherwise called market forces) must form a powerful set of constraints on the activities of dictators. Taxes that are too high will encourage the emergence of black markets, which, themselves, can foster the erosion of political discipline and regard for laws supporting it.

The step from erosion of political discipline to erosion of political support may not be all that large. Spending that's too high, aside from taxation, can erode private saving, reducing capital formation, possibly even in societies in which major items of capital equipment are hand-held metal farming implements and mud-brick private dwellings. Taxes on the wrong items or on the wrong people or groups can be fatal mistakes. Poor planning can lead to endemic queues and the occasional avertable disaster such as a famine. Creation of an extensive administrative bureaucracy can produce a system of corruption (at least what we would call corruption today) in which virtually no public service gets provided without a side-payment (bribe). Systems such as these can erode popular support for the dictatorial ruler. A ruler who, seen from the vantage point of 4000 or 5000 years distance, may appear to have been an absolute dictator might in fact have found it somewhere between advantageous and necessary to consult with groups of wealthy nobles on spending and revenue plans. The Greek city-states of the Archaic and Classical periods certainly lend evidence supporting such interpretations. The extent to which Babylonian and Assyrian kings and the Egyptian pharaohs had to consult on their fiscal policies may be more of an open question than has appeared to have been the case.

The practice of budgetary politics needs some discussion from the perspective of public agencies as well as from that of voters / taxpayers. Government agencies may not be passive recipients of whatever either a monarch, treasury or legislature may be content to assign them. They may engage in a disjointed, decentralized, shoulder-bumping competition for resources involving lobbying, overstatement of accomplishments and understatement of costs, and mandate-imperialism – continuous expansion of original mandates – “mission creep” in contemporary parlance. The supply of public goods and services that emerges from such a system of competitive bidding may bear little resemblance to an optimal supply found in an analytical Lindahl equilibrium, or they may be particularly responsive to different sets of interest groups.

Whether such a funding process results in an aggregate oversupply of public services depends on how closely the revenue-raising process is linked to the demands of agencies. If it is completely de-coupled, there may be little concern on that score.

6.5.3 *Public investment and cost–benefit analysis*

A public good may be provided by purchasing grain from private producers and distributing it free of charge, or at reduced prices, to the poor. Such provision of a merit good is pretty much current consumption. Other public goods, like roads, bridges, and dams, once provided, are expected to yield a flow of services over a lengthy time period. The Lindahl model gave one perspective on how much of both types of public good ought to be supplied but provided little concrete guidance on finding out what the marginal rates of substitution are. It provides more of an analytical benchmark than a planning tool or calculational device. Cost–benefit analysis offers such practical guidance for both current consumption and investment projects, although we will focus more on the investment projects in this subsection.

There generally are many more possible uses of public funds than there are public funds. One task the government has is allocating its funds to the projects that will yield the most public net benefit (that is, benefits minus costs). Although each potential investment project must be evaluated on its own merits, it is always implicitly evaluated relative to other uses of funds. Cost–benefit analysis provides an accounting framework that makes the calculations of net benefits comparable across projects, as well as using sound principles for measuring both benefits and costs in the analysis of each project. This framework answers three major questions: How much to build / supply? How much to spend? When to undertake the project? Subsidiary questions such as the best technologies to use, and what combinations of various inputs to employ are subsidiary to these first three questions, but cost–benefit analysis can assist in thinking about

those issues as well. The general type of answer to the first two of our big three questions is that the last unit of the project (width of road, height of city walls, strength of bridge, and so forth) should add as much in social value as it adds in social cost. This is a variation of the marginal benefit = marginal cost condition that guides virtually all allocation decisions in the economic framework.

In private businesses, the value of an undertaking can be projected with net revenues: prices times outputs, less costs. This excludes the consumer surplus, because a private firm is not concerned with how much surplus its customers derive from its products; it can benefit only from the amount they actually paid, not what they would have been willing to pay but didn't have to. In a public undertaking, it is appropriate to consider consumer surplus as part of the benefit with which the government should concern itself. In general, the benefits that are considered in cost-benefit analysis are consumer surpluses. Frequently it is difficult to observe a demand curve for a public good directly because the good simply isn't available in the private economy. In such cases, the cross-elasticities of demand between the "nonmarket" good and goods that are transacted in markets can be used to get an approximation to the demand curve for the public good.

Let's turn to the temporal structure of benefits and costs, which is where cost-benefit analysis makes some of its most important contributions to evaluation. Suppose we build a public project and expect that the benefits per period will be X for the next n years: $X_0 + X_1 + \dots + X_n$. We've seen that we cannot add benefits from one period to benefits from another period as if there were no difference in their timing. To get an evaluation of the total benefits of the project in comparable terms, we discount the benefits by some discount rate: PDV (present discounted value; sometimes called net present value when the costs are subtracted) $= X_0 + X_1/(1+r) + X_2/(1+r)^2 + \dots + X_n/(1+r)^n = \sum_i X_i/(1+r)^i$. The size of the discount rate will affect the PDV of any given project, as Figure 6.18 shows. In general, the lower the discount rate, the higher

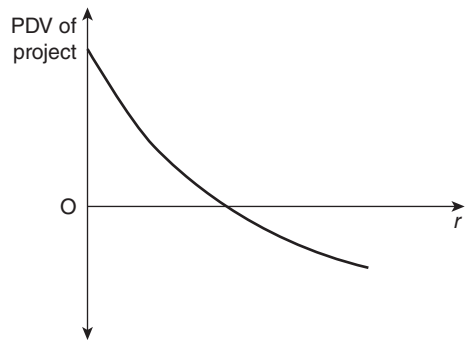


Figure 6.18 Effect of the interest rate on the present discounted value of a capital project.

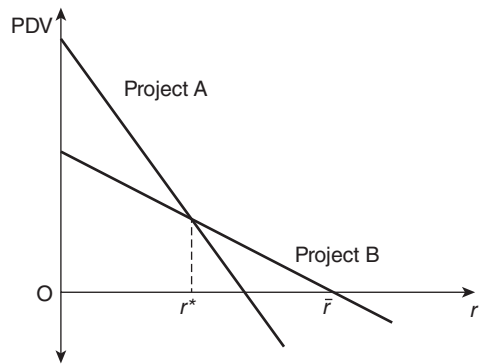


Figure 6.19 Effect of the interest rate on the choice of capital projects.

the PDV. The undiscounted benefit stream need not be constant, so some projects that look very attractive under a high discount rate will look less attractive than other projects when evaluated under a low discount rate. Figure 6.19 shows this comparison. From a discount rate of zero to r^* , project A has the higher PDV, but from r^* to $\bar{r}$, project B has the higher PDV. What can cause these reversals in evaluation? First, project A's benefits stream could dwindle over time while project B's grow. Project B may be a facility in an area whose population is expected to grow over the planning period. The costs may show up much later in project B than in project A and consequently get discounted more heavily than the benefits.

Figure 6.19 also shows two different criteria for evaluating projects: the present discounted value and the internal rate of return (irr), which is the discount rate required to make the present discounted value zero. If we picked between competing projects on the basis of the irr, project B would be selected although project A would deliver a greater PDV over discount rates lower than r^* . The irr is independent of the actual discount rate – either the rate at which the government borrows funds or the rate at which it chooses to discount public benefits. On an irr criterion, both projects warrant undertaking, but if projects A and B are mutually exclusive – if undertaking one of the projects eliminates the conditions for undertaking the other – the preferred project depends on the discount rate.

The use of shadow prices for pricing both inputs and outputs distinguishes cost–benefit analysis from private profitability calculations. As we have noted in passing several times before, the shadow price of an input is the resource cost of that input – the value of foregone output in the economy from the withdrawal of a unit of that resource from its current activities.²⁹ With perfect markets that would equal the market price, but with imperfect markets, missing markets, and market distortions, the values can diverge in either direction. For example, in an economy in which some employment takes place in a “protected” sector that pays higher wages than the rest of the economy, and some workers are willing to accept some period of unemployment to await the possibility of finding employment at the higher wage, the wage in the protected sector is higher than the resource cost of labor. The shadow price of labor would be somewhere between the wage in the protected sector and that in the unprotected sector; if the protected sector were small enough, it might be simply the wage in the unprotected sector. To the extent that wages exceed the shadow price of labor in a protected sector, the market rate of return on capital in that sector would understate the shadow price of capital, since the labor costs are overstated. A public project should be evaluated using a shadow price of labor below its market wage and

a shadow price of capital above its market rate of return.

When considering a project, if we were to look across the anticipated stream of benefits and costs, discounting them both, we could find that delaying some projects would actually increase their PDVs. If the use of a facility would not be expected to begin for several years after completion of the project, it may well make sense to wait for several years to start the project. This is a matter of whether to “build ahead of demand,” and waiting for demand to emerge commonly is the superior choice.

How could we use this analysis to think about ancient construction projects? Suppose the government were able to command *corvée* labor at a zero price to itself (the workers feed themselves while on such duty), but the cost of materials were half the cost of the project (suppose the materials were imported so their resource costs could not be hidden through the exercise of sovereign power). The cost to the government looks like the cost of materials, plus some supervision and coordination costs (palace labor, which, let’s suppose, could do other tasks instead, and get fed besides). Nonetheless, the diversion of the *corvée* labor from, say, agricultural maintenance work (the government is smart enough not to pull them out of private agriculture during the harvest but does not account the value of work done during “slack” periods, such as maintaining dikes, repairing tools, and tending animals and house gardens) has a real cost to the output of the economy. If the government imposes *ad valorem* taxes on that output, it will lose revenue. The (marginal) net benefits of the construction project probably are negative, which simply says that the sum total of the economy’s production, across all activities, is smaller with the project, even including the benefits flowing from it, than without.

Whether ancient construction managers ever thought in terms of discounting future benefits from facilities that were directly productive (for example, irrigation works as contrasted with, say, temples, which are another matter) is an open question, but one for which it might be possible to secure some evidence that might

not be much worse than the evidence on which other bold conclusions have been drawn. For a sufficiently small and distinct project, say an *extension* of an irrigation system (not an entire system), it might be possible to estimate roughly the additional agricultural production permitted. From other information it might be possible to make a similarly rough estimate of the manpower required to build it as well as an estimate of the materials. The cost of manpower could be estimated in any of a number of ways – direct payment records for other activities, comparable or not, with allowance if deemed necessary for differences in skill levels; marginal productivity in agriculture (possibly a more remote possibility; average product would be easier but not useful, but information on land rents could help in getting a rough estimate of MP_L); soldiers' rations, suitably adjusted; and so forth. At the very least, an internal rate of return might be estimable from information along these lines. Evidence from land sale contracts might be persuaded to yield information on private discount rates, and these could be compared with estimated internal rates of return.

An example might help. Let's consider evaluating a project 4000 years ago. Suppose we have a town located on an island in a river. It's existed for several generations and has been thriving lately. To date, however, access between the town and its hinterland has been by ferry. The ferrymen have been providing the town with adequate service, but the ruler of the town and its hinterland thinks about building a bridge. The advantage of the bridge is that people wouldn't have to queue up and wait for a large enough group to engage a ferryman, and the crossing would take less time on the bridge than in the ferry. Additionally, people taking goods in and out of the town wouldn't have to unpack and repack donkey- and vehicle-loads (supposing that the ferry's rowing stock doesn't take animals, or at least loaded animals that might be unstable on the water, and is limited on vehicle capacity it can accommodate). Suppose that, on average, foot travelers save 15 minutes on a one-way crossing. Let's suppose also that the ferry service has been funded by the ruler and that no charge is made for a crossing. The annual

cost reduction for foot travelers crossing the river is the value of 15 minutes per crossing, times the number of crossings per year. The cost reduction for shippers, let's suppose, is a half-hour per crossing, times the number of crossings per year. If the crossing cost had been lower, both categories of ferry passengers would have taken more trips too (that is, the individual and aggregate demand curves for crossings are not vertical).

Figure 6.20 illustrates this view of the river crossing costs. Demand curve D_c is a derived demand based on the values of the various transactions conducted after the crossings, but the cost reductions in which terms we've been speaking are values of time savings, so it is not entirely misplaced to think of the slope of this demand curve as reflecting the value of those time savings. Price line p_{ferry} is the ferry cost per crossing, and p_{bridge} is the bridge cost, per trip, and the demand curve represents the annual demand for crossings. With the ferry c_f crossings are made; with the bridge, we anticipate c_b crossings. The area between the price lines and under the demand curve is the increase in consumer surplus on an annual basis, measured in whatever numeraire might prove useful. Hourly wages probably weren't sufficiently widespread (particularly among the predominantly agricultural population) to make the wage rate a handy measure of the value of time. Chances are that the ruler wouldn't have thought to ask passengers how many handfuls of, say, chickpeas they would have been willing to give to shorten the crossing time and make it

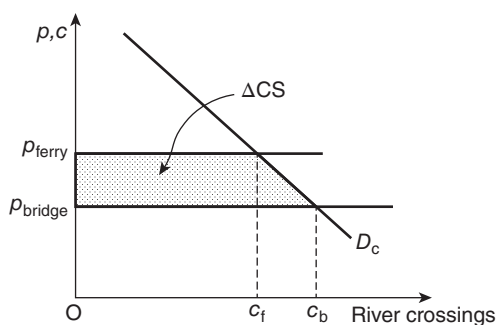


Figure 6.20 Difference in consumer surplus between two capital projects.

drier. (Unfair! We haven't introduced the possibility that the bridge crossing might have been less uncomfortable than the ferry crossing and that passengers might have been willing to pay a bit for the added comfort!) But suppose he did; chances are a non-negative number of handfuls of chickpeas would have been offered.

This consumer surplus, in terms of chickpeas, is our measure of gross benefits per year. Without trying to break it down into handfuls of chickpeas per crossing and numbers of crossings, let's just say that this annual consumer surplus area is equal to 15 000 handfuls of chickpeas. Assume that the population is stable, as is the per capita demand for crossings, so the gross benefits are the same each year the bridge operates. Suppose that the labor and materials costs of building the bridge amount to one hundred thousand handfuls of chickpeas, and the entire cost is paid out in the first six months of year one. For simplicity, assume that there are no maintenance costs during the ensuing years of use. Let's use a 10% discount rate. The gross benefits in the first year are 7500 hfcs (handfuls of chickpeas); the net benefit is negative 85 000 hfcs. We ignore discounting during the first year, beginning it in the second full year of operation. The stream of gross discounted benefits looks like $15\,000/1.1^0 + 15\,000/1.1^1 + 15\,000/1.1^2 + \dots + 15\,000/1.1^{30} + \dots$, which amounts to 159 191 hfcs over 30 years (the PDV isn't quite zero by year 30 but it's down to around 300 hfcs per year, which is negligible). Subtracting the present discounted value of the construction cost yields roughly 52 000 hfcs through 30 years of the bridge's lifetime. The net present value of the benefits stream turns positive in the first few months of the eleventh year.

Let's talk to the devil's advocate for a minute about this example. *"Every city on a river, at least every city that amounted to anything, had a bridge unless the river was just too wide to be bridged with the available technology."* Let's take the first half of the comment first: "every city that amounted to anything..." Does that mean that the cities that "didn't amount to anything" were too small to "afford" a bridge, even if the river could be spanned? Does it mean that when cities were small and young, they used ferries, but when they got large enough to support the traffic, they built a bridge? Let's look at the second qualification: "unless the river was just too wide..." Does

that mean that if the bridge "cost too much" it wouldn't be built? *"Well, how would the ruler have been able to calculate when his city could support a bridge? Isn't that assuming more knowledge of the calculational and accounting techniques than societies had then? How could they estimate demand curves?"* People certainly could calculate, and they were able to approximate quite sophisticated mathematical relationships. Let's think of the calculations as being in two parts – the value of the cost savings and the discounting. Some observations of the efforts involved in ferrying, combined with some introspection about alternatives might yield an approximation to the average time differences. The value of the time difference might emerge more slowly, but again, some observations about what kind (magnitude) of difference a half-hour could make in what someone was able to produce in various endeavors, and the pains people might take occasionally to save themselves some time, could help arrive at some quantitative estimates. For the discounting, surely people routinely made simple, intertemporal choices and had some implicit idea of what they were willing to trade in exchange for particular amounts of some commodities. For the ruler, he would have had alternative things he could have done with the hundred thousand hfcs, and he might have been able to conceptualize when the observable benefits would have emerged. Some pairwise comparisons could have winnowed out something like a discount rate.

Suppose the ruler were to consider a second bridge after a few years of the first one's opening. Would it be worth the cost? If our demand curve for river crossings were unchanged, the fifteen thousand hfcs of consumer surplus would simply be divided between two bridges, at the cost of another 100 000 hfcs. We had only 52 000 hfcs of net present value of benefits from the first bridge. From the perspective of year zero, looking ahead, say five years, the present value of the cost of the second bridge is only about 62 000 hfcs ($100\,000/1.1^5$), but our net present value of benefits after the first bridge is only some 52 000 hfcs, so the additional bridge would not repay itself in terms of its contribution to society's productivity. If we waited five years before thinking about the second bridge, we would have a second "time zero" from which we would begin discounting

benefits. However, the first bridge would absorb half the benefits, so we would be looking at a 100 000-hfc expenditure that could be covered by only about 80 000 hfc NPV of benefits. So the second bridge still would not yield the required net present value criterion. As long as the first bridge didn't wear out, the second bridge would never pay. Of course, when the first bridge had deteriorated to the point where it could handle only a fraction of its original traffic, or required so much in maintenance expenditures to keep it operable, a second bridge could be built, but it really would be just a new, single bridge.

6.6 Regulation of Private Economic Activities

Regulations are general rules or specific actions imposed by government that direct producers to either do things they would not otherwise do or enjoin them from doing things they otherwise would. As such, they interfere with market allocation mechanisms; they intend to. They are a sort of half-way house between nationalization and taxation. It is intuitive to think of regulations as being applied to producers – firms. This is equivalent to inferring that a tax on income from capital taxes the owners of capital, which may or may not be the case, as we have seen. The analogy is not quite parallel though. We should think of regulation as being applied to markets, because restrictions on one side of a market (the producer's) will cause corresponding adjustments on the other side (consumers'). Restrictions on what can be sold are equivalent to restrictions on what consumers can buy. Restrictions on price restrict what producers will be willing to supply. Restrictions on who can own parts of certain types of firms will affect the supply of capital to those types of firms.

There can be many specific targets of regulation. We list a number of popular targets: entry into an industry; price ceilings; price floors; the structure of prices (the level of some prices relative to other prices); profitability; product quality; input choices; technology choices; output choices (production of some goods may be prohibited – for example, prohibition of manufactures in colonies); working conditions; obligation to serve certain potential consumers;

ownership; import quantities; export quantities. Undoubtedly there are others.

Earlier theories of regulation followed what is called the “public interest” approach, on the assumption that the purpose of government regulation of private economic activity is to improve social welfare. Repeated findings that regulated firms frequently or usually benefitted from the regulation suggested another major approach, which could be called the “private interest” approach, sometimes called the economic theory of regulation. There are quite a few specific models in this corpus, the most sophisticated containing both a demand side (the demand for regulation) and a supply side. Comprising the demand side, one way or another in many of the models, is private firms' demand for the protection that cartelization would offer them in circumstances in which antitrust sentiments (and sometimes laws) impede or prohibit direct, private cartelization. Government comprises the supply side since it is government that offers regulation. In legislative models, typically legislators supply regulatory characteristics (prices, say) that maximize their expected voting majority.

Other models in the economics of regulation approach focus on the potential benefits to individuals in government as a prime source of motivation for regulation. Many regulations create rents. For instance, government erection of entry barriers in an industry restricts supply without affecting demand and consequently creates the conditions in which consumers' bids for the available supply of a good or service exceed its factor costs, and – voilà! – we have rents, just like the monopoly rents we found in Chapter 4.

Economies in antiquity were familiar with economic regulation. In Athens (Attica), a prohibition on grain exports is believed to have gone back to Solon in the early sixth century B.C.E (Figueira 1986, 150). While more evidence is published for Athens than most other Greek city states, Bresson (2008, 40–41) reports price controls in Greek cities around the Aegean from the third century B.C.E, through the first century C.E. Osborne (1987, 105–107) reports a regulation from Thasos prohibiting forward sales of grape juice and wine, a regular practice later in Roman Italy. In the Classical Period, Athens implemented a number of regulations,

particularly in its food markets, imposing a constant spread between wholesale and retail prices (it is unclear whether this was absolute or a percentage; and transactions taking place outside the Piraeus or Athens were unregulated: Rosivach 2000, 50), and a prohibition against hoarding of grain (Figueira 1986, 151–152), in addition to inspections for purposes of taxation, laws banning fraudulent practices, and weights and measures regulation (Migeotte 2009, 143–152). Migeotte (1997, 34–43) distinguishes between regulatory practices in ordinary times and exceptional measures, public or private, in distressed times. And then there was the regulation that obliged Athenian citizens engaged in grain trade to bring all the grain they carried to Athens, keeping them from diverting to locations offering higher prices (Figueira 1986, 150; Osborn 1987, 97–104). The Athenian regulation of the spread between the wholesale and retail prices of grains, by not depressing the price offered at port, did not depress supply below what emerged in distressed situations.

Regulation was equally frequent in the Roman Empire. The *lex Iulia de annonae* dealt with speculation and regulation of the grain market (Erdkamp 2005, 265) at a larger geographical scale, while numerous local regulations, particularly regarding the grain markets, proliferated across provinces and locales (Erdkamp 2005, 263). Sometimes the regulations had the theoretically expected effect rather than the effect intended at the time: Erdkamp (2005, 283) reports that the emperor Commodus fixed prices of “all kinds of foodstuffs, only to see the shortage increase.” As an interesting combination of regulations, Erdkamp (2005) cites Lucius Antistius Rusticus’ decree in Antioch in C.E. 93, which combined a price ceiling with mandatory offering of stocks of grain by people holding them. Despite the example (probably long forgotten) of Antistius Rusticus’ dual regulations 150 years before, Julian imposed a maximum price for grain at Antioch in 362/3 C.E., and the market supply dried up, which apparently was the common market response in multiple cases (Erdkamp 2005, 291).

6.6.1 *Rent seeking*

This structure of regulation, particularly with quantity regulations such as quotas, forms the conditions in which private agents have incentives to spend real resources to capture these rents. The maximum they can afford to spend in this effort is, of course, the amount of the rent itself. These costly efforts can include such innocent-appearing activities as polishing applications especially well, maintaining an office in the capital, adding unneeded capacity if licenses are issued on the basis of firms’ relative production capacities. More obvious activities include hiring government regulatory administrators after they retire from government, putting their incompetent nephews on the payroll before they retire, and outright bribery.³⁰ These rents are not additional production but, at the very best, redistribution of income already produced.

The efforts to capture rents from an import quota and the licensing required to implement it raise the domestic price of exportable goods relative to the c.i.f. price of imports (world prices of imports and exports remain unaffected by rent-seeking in any one country) as resources are sucked out of domestic production of exportables and into the domestic handling of imports, an activity greatly expanded by the efforts to capture licenses and the rents created by the licenses. Other regulations can create comparably extensive re-arrangements of resource allocations.

One of the implications of rent-seeking is that a substantial proportion of monopolies in an economy – or oligopoly situations – may be created by governmental regulation of one sort or another. Firms can spend their efforts – real resources – to acquire the access to such monopoly rents, but in doing so can erode part or, in the limit, depending on the extent of the competition for the rents, all of the rents in chasing them. In this view, which is shared by some economists about some situations, the social welfare losses from monopoly are not just the little triangle identifying what consumers would have been willing to buy but the monopolistic seller couldn’t offer without eroding his

revenue. They are the entire rectangle previously identified as monopoly profit. Take a look back at Figure 4.9.

6.6.2 *The costs of regulation: the Averch-Johnson effect*³¹

We can show how the effect of a ceiling on a natural monopoly's rate of return on capital would affect resource allocation at the level of the firm. Suppose our firm has a production function $q = f(K, L)$. It maximizes its profits as $\Pi = p(f(K, L)) \cdot f(K, L) - rK - wL$, subject to $[p \cdot f(K, L) - wL]/K \leq s$. The first term in the objective function really just amounts to revenue = price times quantity, in which the price is a function of the output produced by the firm (remember that we're dealing with a monopolist, which affects the price of its output by its quantity decisions). The rK and wL terms are just subtracting its labor and capital costs. The term s in the constraint is the allowable rate of return, and in the numerator of the left-hand side of the constraint, we have the price-times-quantity (that is, gross revenue) term minus the production costs that are allowable to the regulator; this gives net revenues (or "profits"), divided by the capital stock, which is equivalent to the rate of return the firm earns on its capital. The fact that this quantity has to be less than or equal to the maximum allowable rate of return, s , embodies the rate-of-return regulation.

Recall that the $MRT = MRS$ condition for production efficiency calls for $(\Delta f / \Delta K) / (\Delta f / \Delta L) = r/w$ (the ratio of the marginal product of capital to the marginal product of labor should equal the ratio of the cost of capital – the rental on capital, r – to the cost of labor – the wage rate). All right. Set up the Lagrangean (just imagine it) and take the first-order conditions for the use of capital and labor – that is, adjust the capital and labor used by the firm to make the increments to the Lagrangean go to zero. The first-order condition for capital will give us an expression that the marginal product of capital (MP_K) is equal to something or other and that the MP_L is equal to something else. Let's look at the ratio of these two expressions that come out of the

Lagrangean for the rate-of-return regulated firm: $MP_K / MP_L = (r/w) - [\lambda(s-r)/(1-\lambda)w] < r/w$. Unless the regulator picks the allowable rate of return, s , to be exactly equal to the cost of capital to the firm, it will cause $MRT \neq MRS$. If the regulatory constraint binds the firm (that is, it "bumps up against" the constraint), $\lambda > 0$ and $s > r$, and the firm uses a higher capital-labor ratio than the cost-minimizing ratio. In short, it uses more capital than is socially optimal.

The regulatory agency may have had in mind some monopoly rate of return $s' > s$ when it set the ceiling on the firm's rate of return. Indeed, the output of a firm regulated in this manner will increase as the regulator forces down the value of s . In this sense, some of the intended welfare benefit of the regulation will occur – specifically, the monopolist's tendency to withhold output will be relaxed somewhat – even though there is some cost of achieving that benefit.

The fact that the regulator doesn't know what the regulated firm's cost of capital is exemplifies one of the most stubborn problems involved in administrative regulation of private activity: to make the optimal decisions, the government needs to know a considerable amount of information, and the information-gathering and information-processing expenditures of contemporary regulatory agencies demonstrate just how costly that is. To make the first-best regulatory decisions, in fact, the government needs to know everything the firm does, which means that it virtually has to replicate the entire administrative apparatus of the firm, plus keeping its own staff, an expensive proposition. Some economists believe that the cost of legislation – the administrative costs plus the induced losses – may well exceed the welfare losses from monopoly in the first place (although you will note that this view tacitly accepts a public-interest view of the rationale for the regulation). Even if the regulation does, in fact, improve welfare by restraining monopoly, it does so at an irreducible cost of distortions in marginal conditions. Efforts to apply new regulations to stop the losses from the initial regulations generally will involve what has been called "grasping the tar-baby," and becoming mired in an ever-increasing concatenation

of complicated and costly regulations. Repeated interactions between regulator and regulated firm, as well as the ability to observe different firms, are sources of information to regulators, but making these observations and keeping them current still is a source of cost to the government.

The Averch–Johnson effect is just one example of the many distortions induced by regulations intended (ostensibly) to rectify other distortions. A product-safety law may eliminate from the marketplace (stop production – and eventually consumption – of) products to which are attached certain probabilities of undesirable accidents for the user. Either making the product safer (at higher cost) or eliminating it from the market may seem like an innocuous act, but some consumers may not be able to afford the safer, higher-cost product and have to forego the product altogether; if the product is removed entirely, some of the same consumers may not be able to afford its next-cheapest substitute. Clearly there is some benefit – either to them or to society, possibly both – from their consumption stream having a lower probability of undesirable results associated with them, but those must be balanced against the reduced consumption, especially to the extent that bureaucrats charged with enforcing such regulations have incentives to overprescribe improvements. Consider workplace safety laws. Very laudable indeed, some of them especially, but some workers may have been willing to accept higher wages to work at such employment. The root problem in this situation may be low income, which has been a vastly more difficult problem to remedy than some of its concomitants have been, and the product safety regulations may, after informed public discussion, be a second-best solution to a problem that can't be solved directly. Revising working conditions – rules, technology, and so forth – to increase safety yields a safer work environment, which has a particular value but against which must be debited the losses of the risk compensation in the wage rates of the workers in those occupations. (We will discuss compensating differentials in wages in the chapter on labor.) In the contemporary world, it is in the self-interest of Western workers to demand greater worker safety in industry in the developing world as long as the costs of supplying safer working conditions exceed the risk-premium in workers' wages

(which should be the case, or the work places would have already been made safer – which may strike some as a Panglossian argument, but under competitive conditions among thousands of producers in dozens of countries, probably is not a wide departure from the facts in the case). Clearly this is another subject that may raise social scruples and hackles, and opinions on the subject may be vigorously held. Societies may very well decide that certain standards of work-place safety should not be undercut, and there may well be net benefits to implementing such preferences.

These two examples may seem far afield from the interests of students of the ancient Near East, Egypt, and the Aegean, and, as you may have come to expect, we have some comments on that. Working conditions in particular, and the quality of purchased food, particularly in cities by nonagricultural populations, occasionally come to attention in the ancient records, both written and artifactual. Occasionally, rulers or their agents may have tried – and left evidence in the records – to ameliorate these situations. This analysis identifies some of the issues they would have faced in addressing those issues. It may also go some distance to identifying and accounting for the limits of their efforts.

6.7 The Behavior of Government and Government Agencies

We have touched on some ideas regarding the economic motivations of government – the state – and government agencies in sections 6.5 and 6.6. We collect some of those ideas here and extend them a bit. The motivations of a state in general and state bureaucracies in particular can help predict their economic choices and consequently help explain some of the evidence on past choices. We offer a brief introduction to some of the economic theories of government and government agencies and some concepts regarding the decentralization of government functions and authority.

6.7.1 *Theories of government*

The reason for discussing this topic before this audience is that in thinking about the policy

choices of governments it will prove useful to be able to conceptualize one's beliefs about what government was trying to do, at least as a background condition, when it took the actions whose records or legacy otherwise are left. The economic models of government behavior certainly aren't the only ones possible, but they span pretty widely the possibilities and at the least offer some baselines from which the reader can depart. While some of these models are at least as much borderline philosophies as they are positive models (normative models at best), there has been some effort to develop testable models of government actions along the lines of these theories.

The contractual theories of government, descending from the philosophies of Rousseau and Locke, and represented by Rawls and Nozick among contemporary thinkers, form the basis for social choice theory, with its emphasis on collective decision making, voting, and accommodating individual preferences. These tend toward the normative end of the analytical spectrum, but they fit well with benefit principles of taxation and public sector pricing. To dismiss these theories of government and the economic analysis either deriving from or supporting them on the grounds that most ancient governments tended to fall toward the dictatorial end of the spectrum would be short sighted. The concept of legitimacy from political thought rests on the satisfaction of at least implicit contracts (another subject developed in the economics of labor markets, which we will discuss in a later chapter) and the provision of at least certain minimum levels of benefits. There is more to the economics and politics of dictatorial government than meets the eye, particularly when one thinks in terms of the dictator maximizing an objective function subject to constraints.

The Leviathan view of government, with Hobbes as a major, motivating thinker and represented currently in the economics literature by James Buchanan, sees government as maximizing its own size and power (both of which require definitions and measurement) subject to constraints coming from both the environment and the governed. This view is not inherently incompatible with the concepts deriving from the contractual view. This view of government has supported a number of models of the behavior of administrative agencies

that focus on the utility functions of individual bureaucrats.

Government structure and its behavior are determined by the incentives and constraints of different classes in the Marxian view of government. These models tend to blur the boundaries of normative and positive analysis, although many of the issues brought to the fore by this literature have proven amenable to neoclassical economic modeling.

6.7.2 *Theories of bureaucracy*

Models of bureaucratic agencies begin with a utility function for the agency administrator, focusing on the instruments she has at her disposal to affect her on-the-job utility. These instruments tend to be such variables as budget sizes and size of staff. As we noted earlier in this chapter, pay and promotions tend to be relatively restricted and consequently have little leverage on bureaucratic behavior. Welfare of the governed generally does not appear in the utility functions in these models. However, in models of the behavior of regulatory agencies, particularly those under the influence of the concept that government supplies regulation to industries demanding it, indicators of the wellbeing of the clients are relevant.

One of the facts of bureaucracy that these theories emphasize is that indicators of productivity are less clear-cut than they are in private economic activity and hence less easy to observe. Rulers probably understand this pretty well, and one of their defensive strategies in defending themselves against the growing independence of government agencies and individual bureaucrats is to establish other agencies with overlapping scopes of responsibility – not necessarily identical, but with enough mutual coverage to give the ruler information on relative efficiency and the individual agencies the incentives deriving from competition. These relationships between rulers and bureaucrats involve what are called principal-agent relationships, which we will discuss in more detail in the chapter on the economics of information. Briefly, a principal is the individual who establishes goals and makes the rules and the agent is the individual who is charged with following the rules and seeking to reach the goals, subject to varying degrees

of observability by the principal. The principal would be unwise to take the statements of the agent(s) as necessarily truthful. The two have different sets of information, and each can dissemble to the other. A wise principal will find incentives to offer the agent that elicit efforts that produce the results the principal desires; generally this involves finding common incentives, which may be a combination of rewards and punishments, with rewards typically being more effective.

6.7.3 *Levels of government*

Decentralization of government takes a number of forms. A highly centralized government of a large territory will find it useful – imperative – to operate offices of agencies at locations outside the capital. Many government operations are directed at the conditions of specific locations and require information from those locations. The major shots may still be called from the center, but not all the contingencies that will arise can be spelled out in standing instructions. Consequently, decentralized bureaucrats will have some type of residual authority which, depending on the duty area, may be quite extensive. In these cases, the principal-agent problem is back in full force, with the added informational cost of distance.³²

Another form of decentralization of government is federalism – a layer of separate governments with varying degrees of authority in progressively more spatially limited jurisdictions. Whether the decentralization actually involves any semblances of separate sovereignty or not (in many of the ancient states of the Near East, it may not have), divisions of authority among levels of government involve decisions about what authorities to delegate. One of the principles of the delegation of authority that has emerged from the economics of federalism is the benefit of having territorial correspondence between taxing authority and responsibility for provision of public goods and services. A contemporary example from the United States may illuminate the principle quite well: the provision of public welfare benefits. If states or lower levels of government have the principal responsibility for providing public support for the unemployed and others considered to qualify for public income assistance, but the people qualifying for

such assistance are mobile among states, those individuals will have the incentive to locate so as to maximize their net benefits (net of movement costs). If states have the responsibility to provide the revenue as well as the authority to define qualifications and benefit levels, there will be no redistributions of tax revenue among states but there may be incentives for each state to ratchet up their qualifications and ratchet down their benefit levels to force potential recipients into other jurisdictions. Definition of qualifications and benefits, as well as extraction of revenue, at the largest spatial scale over which potential recipients might move (the nation?) would eliminate the incentives for both migration for benefits and shifting of burdens across jurisdictional lines. (Similar principles govern the spatial definition of optimal currency areas, which falls under the purview of the chapter on international economics.)

Examples of good candidates for local provision and local revenue raising are water supply, education, and libraries. There are relatively well defined – and small – spatial ranges over which consumers can travel to obtain these services. This type of public service provision has been formalized in the Tiebout (1956) model of local public government, in which mobile individuals band together and provide themselves the array of local public services they demand through taxation, with membership in the community as the admission ticket. These goods and services are not “pure” public goods such as national defense, but are characterized by congestability and at least partial excludability, although their private provision would still be subject to serious pricing problems and suboptimal supply.

6.8 **Suggestions for Using the Material of this Chapter**

Ancient taxes will probably be unobservable archaeologically but there is surely evidence of them in texts. Ancient historians and philologists may want to keep in mind putative demand elasticities of various goods that were taxed at different rates. To what extent did ancient fiscal authorities implicitly understand the shiftability of a tax base? Did they tax goods that we would think were in less elastic demand more?

A large proportion of architectural remains, especially in cities, but also in the form of rural temples, were public goods – decisions about how much of them to build were made differently from decisions about private goods. Even in smaller towns, clay water pipes along the sides of streets indicate some public good; some organization larger than the household decided to extract resources from people all along the street to install the drainage pipes. Similarly, raised central ribs in cobbled or paved streets in a town indicate that the decisions regarding the street construction were made at levels higher than the households fronting on the street. We could envision something like an ancient block party deciding on such a construction but I doubt we could convince ourselves that the decisions were really made that way. These simple structures imply a collective decision making, with no

aspects of contemporary democracy implied. A king's or town mayor's advisors might well make a recommendation that the local autocrat then implements – with resources extracted from the subjects, probably no less willingly than we part with tax payments today.

Irrigation systems may have been public goods to a considerable extent – or mixed public-private goods. Since takeoff of water at one point affects water availability at points downstream, all potential users along a stretch of waterway would have had an incentive to contribute to the infrastructure in order to have a voice in the water allocations. This set of incentives would have at least made the infrastructure a club good if not a pure public good. I cannot predict the archaeological correlates of this decision structure, although some may be found in tablets in Mesopotamia.

References

- Arrow, Kenneth J. 1963. *Social Choice and Individual Values*, 2nd edn. New York: John Wiley & Sons, Ltd.
- Atkinson, Anthony B., and Joseph E. Stiglitz. 1980. *Lectures on Public Economics*. New York: McGraw-Hill.
- Averch, Harvey, and L. L. Johnson. 1962. "Behavior of the Firm under Regulatory Constraint." *American Economic Review* 52: 1052–1069.
- Ballard, Charles L., John B. Shoven, and John Whalley. 1985. "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States." *American Economic Review* 75: 128–138.
- Breasted, James Henry, editor and translator. 1906. *Ancient Records of Egypt. Historical Documents from the Earliest Times to the Persian Conquest. Volume II. The Eighteenth Dynasty*. Chicago IL: University of Chicago Press.
- Bresson, Alain. 2008. *L'économie de la Grèce des cites. II. Les espaces de l'échange*. Paris: Armand Colin.
- Christ, Matthew R. 2002. "Liturgy Avoidance and Antidosis in Classical Athens." *Transactions of the American Philological Association* 120: 147–169.
- Courtois, J.C. 1982. "L'Activité Métallurgique et Les Bronzes d'Enkomi au Bronze Récent (1650–1100 avant J.C.)." In *Early Metallurgy in Cyprus 4000–500 BC*, edited by James D. Muhly, Robert Maddin, and Vassos Karageorghis. Nicosia: Pierides Foundation, pp. 155–175.
- Davies, J.K. 1967. "Demosthenes on Liturgies: A Note." *Journal of Hellenic Studies* 87: 33–40.
- Davies, J.K. 1971. *Athenian Propertied Families, 600–300 BC*. Oxford: Clarendon Press.
- Dikaïos, Porphyrios. 1969. Enkomi: Excavations 1948–1958. Volume I: *The Architectural Remains. The Tombs*. Mainz am Rhein: Philipp von Zabern.
- Erdkamp, Paul. 2005. *The Grain Market in the Roman Empire: A Social, Political and Economic Study*. Cambridge: Cambridge University Press.
- Figueira, Thomas. 1986. "'Sitopolai' and 'Sitophylakes' in Lysias' 'Against the Graindealers': Governmental Intervention in the Athenian Economy." *Phoenix* 40: 149–171.
- Fisher, Irving. 1930. *The Theory of Interest; As Determined by Impatience to Spend Income and Opportunity to Invest It*. New York: Macmillan.
- Garnsey, Peter. 1988. *Famine and Food Supply in the Graeco-Roman World; Responses to Risk and Crises*. Cambridge: Cambridge University Press.
- Hammond, Peter J. 1990. "Theoretical Progress in Public Economics: A Provocative Assessment." *Oxford Economic Papers* 42: 6–33.
- Harberger, Arnold C. 1962. "The Incidence of the Corporation Income Tax." *Journal of Political Economy* 70: 215–240.
- Harper-Collins-Sansoni. 1988. *English-Italian Italian-English Dictionary*, 3rd edn. Florence: Sansoni.
- Hausman, Jerry A. 1985. "Taxes and Labor Supply." In *Handbook of Public Economics*, Vol. 1, edited by Alan J. Auerbach and Martin Feldstein. Amsterdam: North-Holland, pp. 213–263.
- Heal, G.M. 1969. "Planning without Prices." *Review of Economic Studies* 36: 347–362.

- Holleran, Claire. 2012. *Shopping in Ancient Rome; The Retail Trade in the Late Republic and the Principate*. Oxford: Oxford University Press.
- Jones, A.H.M., and P.J. Rhodes. 1996. "Liturgy." In *The Oxford Classical Dictionary*, 3rd edn. Oxford: Oxford University Press, p. 875.
- Kanawati, Naguib. 1980. *Governmental Reforms in Old Kingdom Egypt*. Warminster: Aris & Phillips.
- Keith, Kathryn. 2003. "The Spatial Patterns of Everyday Life in Old Babylonian Neighborhoods." In *The Social Construction of Ancient Cities*, edited by Monica L. Smith. Washington, D.C.: Smithsonian Books, pp. 56–80.
- Killingsworth, Mark R. 1983. *Labor Supply*. Cambridge: Cambridge University Press.
- Krueger, Anne O. 1974. "The Political Economy of the Rent-Seeking Society." *American Economic Review* 64: 291–303.
- Laffont, Jean-Jacques. 1988. *Fundamentals of Public Economics*. Translated by John P. Bonin and Hélène Bonin. Cambridge, MA: MIT Press.
- Lange, Oskar. 1936. "On the Economic Theory of Socialism: Part I." *Review of Economic Studies* 4: 53–71.
- Lange, Oskar. 1936. "On the Economic Theory of Socialism: Part II." *Review of Economic Studies* 4: 23–142.
- Lippincott, Benjamin E., ed. 1939. *On the Economic Theory of Socialism*. Minneapolis MN: University of Minnesota Press.
- Lipsey, Richard G., and Kelvin J. Lancaster. 1956–1957. "The General Theory of the Second Best." *Review of Economic Studies* 24: 11–32.
- McLure, Charles E., Jr. 1975. "General Equilibrium Incidence Analysis; The Harberger Model after Ten Years." *Journal of Public Economics* 4: 125–161.
- Meade, James E. 1952. "External Economies and Diseconomies in a Competitive Situation." *Economic Journal* 62: 54–67.
- Migeotte, Léopold. 1997. "Le contrôle des prix dans les cités grecques." In *Entretiens d'archéologie et d'histoire, 3: Prix et formation des prix dans les économies antiques*, edited by J. Andreau, P. Briant, and R. Descat. Toulouse: St.-Bernard-de-Comminges, pp. 33–52.
- Migeotte, Léopold. 2009. *The Economy of the Greek Cities; From the Archaic Period to the Early Roman Empire*. Translated by Janet Lloyd. Berkeley CA: University of California Press.
- Müller, Christel. 2011. "Autopsy of a Crisis: Wealth, Protopogenes, and the City of Olbia in c.200 BC." In *The Economies of Hellenistic Societies, Third to First Centuries BC*, edited by Zosia H. Archibald, John K. Davies, and Vincent Gabrielsen. Oxford: Oxford University Press, p. 324–344.
- Osborne, Robin. 1987. *Classical Landscape with Figures; The Ancient Greek City and its Countryside*. Dobbs Ferry NY: Sheridan House.
- Postgate, J.N. 1974. *Taxation and Conscription in the Assyrian Empire*, Studia Pohl: Series Maior 3. Rome: Biblical Institute Press.
- Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *Economic Journal* 37: 47–61.
- Rosivach, Victor J. 2000. "Some Economic Aspects of the Fourth-Century Athenian Market in Grain." *Chiron* 30: 31–64.
- Sandmo, Agnar. 1976. "Optimal Taxation; An Introduction to the Literature." *Journal of Public Economics* 6: 37–54.
- Sandmo, Agnar. 1985. "The Effects of Taxation on Savings and Risk Taking." In *Handbook of Public Economics*, Vol. 1, edited by Alan J. Auerbach and Martin Feldstein, 265–311. Amsterdam: North-Holland.
- Shelmerdine, Cynthia W. 1973. "The Pylos Ma Tablets Reconsidered." *American Journal of Archaeology* 77: 262–278.
- Shelmerdine, Cynthia W. 1989. "Mycenaean Taxation." In *Studia Mycenaea* (1988), *Živa Antika* Monograph No. 7, edited by T.G. Palaima, C.W. Shelmerdine, and P.H. Ilievski. Skopje: Dept. of Classical Philology of the University of Skopje; and Austin: Program in Aegean Scripts and Prehistory of the University of Texas at Austin, pp. 125–148.
- Spulber, Daniel F. 1989. *Regulation and Markets*. Cambridge MA: MIT Press.
- Starrett, David A. 1988. *Foundations of Public Economics*. Cambridge: Cambridge University Press.
- Stech, Tamara. 1982. "Urban Metallurgy in Late Bronze Age Cyprus." In *Early Metallurgy in Cyprus, 4000–500 BC*, edited by James D. Muhly, Robert Maddin, and Vassos Karageorghis. Nicosia: Department of Antiquities, Cyprus, pp. 105–116.
- Stern, Nicholas. 1987. "The Theory of Optimal Commodity and Income Taxation: An Introduction." In *The Theory of Taxation for Developing Countries*, edited by David Newbery and Nicholas Stern. New York: Oxford University Press, pp. 22–59.
- The Compact Edition of the Oxford English Dictionary*, Vol. 1. 1971. Oxford: Oxford University Press.
- Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 63: 416–424.
- Tylcote, Ronald F. 1982. "The Late Bronze Age: Copper and Bronze Metallurgy at Enkomi and Kition." In *Early Metallurgy in Cyprus, 4000–500 BC*, edited by James D. Muhly, Robert Maddin, and Vassos Karageorghis. Nicosia: Department of Antiquities, Cyprus, pp. 81–103.
- Wyatt, William F., Jr. 1962. "The Ma Tablets from Pylos." *American Journal of Archaeology* 66: 21–41.
- Zuiderhoek, Arjan. 2009. *The Politics of Munificence in the Roman Empire; Citizens, Elites and Benefactors in Asia Minor*. Cambridge: Cambridge University Press.

Suggested Readings

- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River, NJ: Macmillan. Chapter 18.
- Singer, Neil M. 1976. *Public Microeconomics; An Introduction to Government Finance*, 2nd edn. Boston: Little, Brown.

Notes

- 1 Government revenue from direct government productive activities corresponds to the “wealth finance” concept used in anthropology and which has been taken up by some archaeologists, and revenue from taxation, particularly if it were in kind, corresponds to the “staple finance” concept.
- 2 Actually, if there are cross-price elasticities of demand between a public good and some private goods, some reasonably decent estimates of the demand for the public good can be obtained. This technique has been used in public investment evaluation for goods as diverse as expansion of public roads and improvement of recreational fisheries.
- 3 Of course, in some cases, possibly many cases, even persuading beneficiaries that they are indeed beneficiaries may push the limits of a society’s rhetorical powers, including its science and its laws.
- 4 Marginal costs are easy to measure here – the money price of pipe tobacco – but the marginal benefits require a bit more explanation. We measure them as ratios of marginal utilities, or as a marginal rate of substitution between the pipe smoking and some numeraire good, which puts the utility measures in some real terms, which could be measured in money terms if one so desired.
- 5 I put “victim” in quotation marks to indicate the unusual character of victimization in externalities. In the example of B’s pipe smoking, B could as easily be victimized as could A. If we as analysts have some ethical qualms about pipe smoking in particular or smoking or the use of mood-altering substances in general, that is external to the externality! It is a different issue entirely from the matter of one person’s self-directed action affecting another person’s equally self-directed assessment of personal wellbeing. The models of utility theory are not well suited for the analysis of interpersonal moral outrage except as they are willing to express their outrage by foregoing consumption of something else. This disinterestedness of economic theory not infrequently causes discomfort in observers who are intensely interested in “getting the right answer” out of an analysis. “Getting the right answer,” however, is generally a matter of politics (not political science, but practical politics).
- 6 If a downward-sloping marginal damages curve cut the marginal benefit curve from above, we would have a situation in which small amounts of the externality were a problem, in the sense that they imposed marginal costs greater than the marginal benefit to the direct consumers of the good generating the externality, but if the volume of the activity became large enough, the damages or disutility would disappear. Casual introspection fails to turn up any important externalities with this characteristic. Thus, while the slope, and generally the shape, of the marginal damages curve may be largely an empirical matter rather than a theoretical one, a non-negative slope seems reasonable.
- 7 Losses through inefficiencies involved in the tax-and-transfer system. These will become clearer with the discussions in the following subsections.
- 8 The example of the welfare weights comes from Stern (1987, 47–48).
- 9 On the Pylos Ma Tablets, see Wyatt (1962) and Sheldermine (1973; 1989 and the references cited therein). On Assyrian taxation, see Postgate (1974).
- 10 “At the margin” means that the entire French government revenue need not have cost half its value in lost output, but that expansions of revenue from current levels could be expected, with conservative estimation, to displace as much as half of the marginal revenue collection. See Laffont (1988, 186).
- 11 That is, results of the sort, “X goes up when y goes down, depending on the relationship between w and z ,” derived from graphical or mathematical analysis; contrasted to empirical or numerical simulation results.
- 12 Recall, however, that the household production model does give a number of insights into the allocation of labor between different activities which could be subdivided so as to yield market and nonmarket allocations and allocations

- between activities we could feel comfortable calling “leisure” and nonleisure, despite reservations such as those Becker has expressed about the ill-defined character of the common uses of the term “leisure.”
- 13 Virtual incomes are created by a tax system, which gives individuals the incentive to behave as if they possessed property income different from their actual property income.
 - 14 This model was used in the intertemporal choice section of Chapter 3, but that treatment did not emphasize savings.
 - 15 In the Slutsky equation for second-period consumption, both effects work in the same direction for net lenders.
 - 16 See Atkinson and Stiglitz (1980, 77–78) for the development of the expression in the two-period Fisher model.
 - 17 Stochastic means that the variable so described has elements of randomness in its value. The actual value of a stochastic variable is determined at least in part by random draws whose values are governed by a probability distribution. The “expected value” is what is called “the average” in lay terms. The variance of a random variable (stochastic variable) is its second moment (the expected value or average is its first moment). The variance describes the tendency for different observations of the random variable to cluster about the mean. A high variance means that there is considerable variability about the mean – although the mean is the best predictor of the value of any observation, it is not an especially good predictor when the variance is high. The third moment describes the degree of skewness of the distribution – the degree to which many observations tend to cluster around either high or low values of the distribution rather than the mean.
 - 18 These models are developed by Atkinson and Stiglitz (1980, 99–124) and Sandmo (1985, 293–300).
 - 19 Calculation made by Sandmo (1985, 297).
 - 20 This formulation assumes that any taxes on final products that are used as inputs to other consumer products have the taxes on them rebated. Without this assumption, the expression would be a bit more complicated.
 - 21 This probably seems cumbersome. It is. To get around this type of cumbersomeness, matrices and matrix algebra are used routinely to work with systems such as this.
 - 22 Another term from the lexicon of trade theory, meaning exactly what it says: commodities that can be traded; some things can’t be easily traded, such as housing or personal services that must be delivered in person, such as doctoring or barbering.
 - 23 Ramsey was a brilliant young mathematician at Cambridge University whose attention to two important and difficult economic problems – taxation and economic growth – was attracted by J.M. Keynes and other economists at Cambridge in the mid-1920s. Ramsey died in 1930 following surgery. The taxation article has also formed the theoretical backbone of public enterprise pricing theory, as we will discuss in section 6.5.
 - 24 Reprinted in Lippincott (1939, 55–142). Most centrally planned allocation systems have used quantity directives rather than inferred or created “prices.” Much of the economic theory of central planning developed in the West since World War II has used mathematical programming techniques to determine “shadow prices,” or resource-cost estimates that central planners can use to send signals to decentralized production sectors. Heal (1969) supplements quantity information with factor marginal productivity information at desired output quantities, which is quite close to the information contained in shadow prices. (The Lagrange multipliers on the constraints turn out to be the shadow prices of inputs, since they measure the value of increasing the value of the constraint just a bit – which is exactly the marginal information contained in efficient prices.)
 - 25 For which external trade is not a realistic solution because of the erratic and unpredictable character of the excesses, which could be balanced with alternative shortfalls.
 - 26 “Endogenous” means that the value of a particular variable is determined within, or by, the model in question. This is contrasted to “exogenous,” which means that the value of the variable is determined outside the model (by the analyst).
 - 27 Named for William Vickrey, who discovered the mechanism in 1961, won the Nobel Prize in economics for it in 1996, and died at a rest stop on the Palisades Parkway before he could receive it. His wife accepted it for him.
 - 28 The term appears in *The Compact Edition of the Oxford English Dictionary*, Vol. 1 (1971, 1654) (equivalent to p. 404 of the L volume of the full-size edition), which is reprinted from the edition revised in 1933. The nuances reported there show that the term is well understood, although it may travel under different names in different Western European (and, of course, other) countries. It appears in the Harper-Collins-Sansoni *English-Italian Italian-English Dictionary*, 3rd edn. (Harper-Collins-Sansoni 1988, 612), with

an expression that seems to capture its American meaning satisfactorily.

- 29 An alternative definition of a shadow price is the value of the change in a social welfare function caused by a unit change in public production, times the change in public production per unit change in a particular input.
- 30 Quantitatively, such rents can be important. Krueger (1974, 294) estimated that the government of India, through its licensing of imports, certain controlled commodities, credit, and in the railway industry, created rents amounting to

7.3% of the country's national income in 1964, an amount large relative to the size of national savings. She estimated that the rents from import licenses alone accounted for over 15% of Turkey's national income in 1968.

- 31 Named for Harvey Averch and Leland Johnson (Averch and Johnson 1962), who first analyzed this regulation-induced "over-capitalization."
- 32 See, for example, Kanawati (1980) for an assessment of the spatial dispersal of government offices in Old Kingdom Egypt.

The Economics of Information and Risk

This chapter is relatively detailed. It contains a fair bit of symbolic representation, although if the reader will stay with us long enough to look, it will be found that the mathematics is once again limited to the four basic arithmetic operations. The symbols save a considerable amount of space as well as permit – even force – an enhanced degree of precision in the specification of concepts. Risk is an intricate subject, yet it is common to find explanations of artifactual remains or epigraphic evidence as attributable to, or responses to, risk – with more or less a wave of a hand. The patience for more thorough-going investigations of what might be implied by a “risk” explanation may indeed be limited, but we believe that benefit is to be gained by keeping one’s feet to the fire, so to speak, in being more explicit. This chapter offers the explicit material with which to think more thoroughly and more precisely about risk and its counterpart, information. Some of the results will appear counterintuitive, although we strive to develop the intuitive understanding of them.

7.1 Risk

This section introduces a number of basic concepts that will be used time and again in the study

of behavior under uncertainty and behavior regarding information. We will see that uncertainty and information are something close to opposites of each other. Both involve costs, and there are demands and supplies for each if we cast the problems at hand in such a framework. We begin the chapter with risk because it is intuitively recognized as a pervasive feature of ancient life – especially among scholars who interpret most ancient individual and social actions and organizations as methods of protecting against risk (otherwise why spend so much effort guarding against it?). We will offer methods we believe will help contemporary scholars of the ancient Mediterranean and Aegean to formulate their inquiries about ancient behavior. We also believe that these methods will give insights into the motivations and actions of ancient decision makers – otherwise there would be little for contemporary scholars to gain in using the methods themselves.

We open with a descriptive discussion of the kinds of situations that involve decision making under uncertainty. Following that, we develop some concepts that help us measure risk and think about its relationships to, and effects on, behavior. The third subsection draws together these concepts into a somewhat more extensive treatment of the expected utility model which

crystallizes the kinds of configurations of uncertain outcomes a typical consumer would strive to achieve through various actions. The final subsection delves into the substantive content of the concept of probability, particularly as it applies to the knowledge base of ancient people.

7.1.1 *The ubiquity of risky decisions*

Risk and uncertainty are quite intuitive concepts. We all find ourselves, literally daily, in situations where we may take one of several actions but we don't have all the information to be absolutely sure that any one of the actions will yield us the greatest satisfaction. Shall I take route A or route B to work today? (What will the traffic be like?) Shall I, or shall I not, see the movie X tonight? (Will I like it? Would I agree with the reviews I've read? Will I agree with the opinions of my friends who have seen it already?) Will these apples be as good as they look? (We've all gotten mealy or tasteless apples, and most of us have devised little signal systems to advise ourselves on the tastiness of apples: for any particular variety of apple, its color relative to some standard we've conceived; its firmness, its weight relative to its size; its very size; its smell. But even still, with all these indicators on which we can get information – short of biting into it before we buy it – we might get an apple we would prefer not to have bought.) Will it rain today or not? (Should I schlep my umbrella with me, at the risk of forgetting it somewhere if it doesn't rain – and besides, it's a bother to carry – or am I likely to be far from shelter and likely to get thoroughly soaked and mess up my wool suit before I can find a dry spot? Is it cloudy or clear this morning? If it's cloudy, what inferences do I make about the likelihood of rain this afternoon? What inferences if it's clear? Do I combine "cloudy or clear" with other indicators to make an inference about rain? If I make an inference that it might rain, do I automatically take the umbrella or do I take a chance without it? What information about my daily itinerary, or other items, do I use to make that decision?)

On larger scale decisions – those whose opportunities appear less often and which require more substantial commitments of resources on our part – similar types of issues exist and capture our attention even more pointedly. Shall I take this job or that one? (I may know several important

bits of information about each – remuneration, hours, some aspects of working conditions; duration or term; location; and probably other important characteristics – but about others I simply cannot know now. Will I find my colleagues compatible? How will my remuneration progress? If I am guaranteed certain fixed raises for the next, say, n periods, what will the progress of prices – inflation – do to my purchasing power, and will my employer be willing to recontract if inflation really accelerates or will she make me stick by the original agreement and take a reduction in real remuneration? Will business conditions permit the company to thrive, letting my remuneration rise accordingly and proffering job security, or will they sink the company, leaving me unemployed? Do I believe all that my potential employers tell me? Can I tell for certain that they are being honest and forthcoming, or should I suspect that they are strategically withholding some information they have? Conversely, am I prepared to tell them everything I know about myself as a potential employee, or do I want to behave strategically with them?) How much insurance should I buy, and what kind(s)? Which stocks and bonds should I buy, and in what proportions? Who should I employ for various tasks that I may want to have someone else do, in either private or business life?

In the examples of daily uncertainties, the possible downside risks may be minor inconveniences or minor regrets about resource allocation decisions. We might be able to get even better information about the conditioning circumstances than we start out with, but how much effort is the decision worth? We *could* take one particularly nice looking apple to the cashier, buy it, eat it as a test, and then decide whether or not to buy any other apples that look like it. Frankly, this sounds like a lot of work for a few apples. And even if the one we buy is excellent, does that guarantee that all the others that look like it will be equally good? Do we suspect that we've probably found the very best apple – it's as good as apples get – so that any others we would buy could only be as good or worse, but not better? (We probably have a semi-conscious idea of these relationships when we first look at the bin of apples and process these questions pretty quickly.) How much trouble is it worth to improve the information we have before we

make decisions about seeing a movie or taking an umbrella? How good would the extra information be if we decided to seek it, and how much would it cost us, in terms of either effort or cash? Is there *anybody* who *really* knows whether it's going to rain or not this afternoon?

The value of the resources riding on the "big" decisions commonly makes it worthwhile to collect more information before committing ourselves. Do I think a potential employee is competent? Will he be industrious, if he is competent? How much can I find out about competence or industriousness, and how can I find it out? Can I buy information? Can I induce the people themselves to reveal truthful information? How likely am I to get false information, either through purchase or from the person himself or herself? How much should I spend to learn more, and when should I quit spending and just make a decision?

Suppose we're business people with an opportunity to invest in some extra productive capacity or in a new line of activity. Most likely we have some business competitors. How will they react (what will their reactions do to prices and our profits)? Why do we expect any particular reaction? Is there anything we can do to improve our information before making a decision one way or the other about the investment? Can we be absolutely sure about what our competitors will do?

Put ourselves in a farmer's shoes for a minute. We have to commit resources to seed, fertilizer, equipment use, and considerable effort, as well as the exclusion of alternative cropping decisions, prior to knowing what the weather will be like at critical junctures during the crop season, what other suppliers of the same and competing crops will do, and what the price will be at the end of the season (not to mention subsequent prices if we can store our harvested crop if prices are low right after harvest). There's literally no one of whom to ask these questions. And we could rest assured that if we asked certain people, they'd lie (we might or might not be able to predict the direction of the lie). If we've collected all the information we think it's worthwhile to collect,

are there some productive actions we could take that would reduce the effect of untoward events on our income at the end of the season? Plant several crops? Take a part-time job in town and hire somebody to help with the crops or just forego some amount of planting? Is there any insurance we can buy that doesn't cost more than it's worth to us? Whatever we do, there is some risk, and the part of the risk we've gotten rid of one way or another has cost us something to eliminate (or reduce).

These examples involve just about every topic that we will address in this chapter. We will clarify the logical structure of these "problems" and endeavor to separate issues that are closely involved with one another. We have dealt in earlier chapters with many of the decisions considered here, but in those treatments we assumed, either implicitly or explicitly, that all the information relevant to the decisions was available and was accurate. And, except in the section of Chapter 4 on oligopoly, we assumed that no one believed that any of his actions would affect anybody but himself or herself, but in some of the examples just given, people were not unreasonable to suspect that strategic behavior was in their best interest and that others with whom they were dealing might behave strategically as well. Those models are not so much "wrong" as they are simplifications. They yield considerable insight into the structure of interactions and decisions despite – or possibly because of – their simplifications, and the analysis of this chapter builds on their structure, but with allowance for greater realism. This greater realism comes at a cost, however, which reveals itself in increased complexity and, in general, the restriction of interpretable models to small numbers of interacting agents, frequently only two.

The risks involved in the examples above each have had one of two sources: what is called "the state of nature," or just plain "nature," and the behavior of other agents. These two sources of risk have been called "event risk" and "market risk," but "nature risk" and "behavioral risk" would be equally suitable as descriptors of these two categories of risk sources.

7.1.2 *Concepts and measurement*

So far, we have used the terms “risk,” “uncertainty,” and “information” with their intuitive, colloquial meanings, and that has served us adequately to formulate examples with a reasonable amount of resonance. To study the logical structure of these concepts and the decisions made that involve them, we need a bit more formality, which we attempt to make as painless as possible. To motivate the following discussion, let’s ask how we would *measure* risk, about which we have talked so loosely, implicitly comparing situations of more or less risk. Suppose we consider the price of some good, measured in shekels. From long observation, we can state that 95% of the time, that price varies within a range of two shekels. Stated alternatively, we have made enough observations to calculate what we believe to be the average price of that good, and “most of the time” the price is within a shekel either side of that average. Is that a lot of variability or a little? Without identifying the average price, the variation doesn’t tell us much. If the average price is two shekels, a variation of a shekel on either side of that figure is 50% of the “average,” which is quite a bit for many goods. If the average is one hundred shekels, the range of variability is only one percent, which in many cases would be quite small. We started this paragraph talking about risk and now we find ourselves threatening to become mired in the variability of prices. Perhaps we should expand a bit on the connection.

We picked a price as the variable whose value at any particular time might take any of a number of only imperfectly predictable values. We could have picked rainfall, temperature, labor effort of an employee, heads or tails on a tossed coin, the likelihood of dying from any of a number of causes within the coming year, the ratio of height to weight of 12-year-old children, and so forth. Two factors combine to give these variables risk implications: the extent of their unpredictability, and someone’s making a resource-allocation decision, the outcome of which depends on the particular value of the variable in question. In our list of such variables, rainfall, temperature,

and employee labor effort very well may form important elements in someone’s resource allocation decision. A farmer stands to prosper or go hungry, possibly starve, depending on what values the variables rainfall and temperature take during his planning period (the crop season). An employer may prosper or not fare nearly so well depending on how diligently his employees work (the employer may be a small farmer and the employee may be a part-time hired hand). The probability of dying during the coming year may be an important variable in resource-allocation decisions: if you know you’ll die during the coming year, you won’t buy tickets for the following year’s World Series (or World Cup, depending on your preferences); you might decide to spend your assets at a faster rate than you otherwise would, or you might hoard them so you can leave a bequest to your survivors. If you aren’t sure whether you’ll live or die during the coming year, your behavior will likely be conditioned by your beliefs regarding your prospects. At any rate, some important economic decisions probably ride on the probability you assess to this variable (a “binary” variable – it takes a 1-or-0, yes-or-no, value, as contrasted with a “continuous” variable, which can take literally every value between some lowest and highest values, or a “discrete” variable, which can take only specific values, frequently integers – 1, 2, 3, and so forth – but not necessarily restricted to integers; binary variables are a special case of discrete variables). A probability, as a quantitative measure, can take values between zero and plus one; a probability cannot be negative, nor can it be greater than one.

It would require a stretch to imagine a situation in which the ratio of height to weight of 12-year-olds entered an economic agent’s profit function,¹ but it is not particularly difficult when the variability in that variable might be of interest to a scholar studying economic conditions. Suppose you’re trying to estimate the likelihood of children’s contracting certain diseases, for which the age- (and sex-) specific height-to-weight ratio is a good predictor of predisposition or susceptibility. There may not (in fact probably won’t be) a one-to-one relationship between

the height-to-weight ratio and contracting the disease; some children with a particular ratio will contract the disease and others won't. The coin-toss example has little risk implication unless you're betting on coin tosses, which itself requires either low risk aversion, dishonesty in one person and gullibility in the other, or considerable boredom.²

With these examples of variables in hand, let's proceed to the building blocks of risk. The first is the concept of the random variable. We'll treat "variable" as a primitive, a term that does not require definition, but let's discuss it a bit anyway. A variable is any concept that you want to think about clearly enough to bother defining; you may or may not want to measure it. For instance, the concept of utility involves a variable, but not one we particularly want to try to measure. The concept, when sufficiently well defined, guides our thinking about many other variables (concepts) that we *would* want to measure, but our understanding of the utility concept tells us that we don't want to try to measure it (we might, however, want to measure ratios of marginal utilities). A random variable (sometimes called a stochastic variable) is a variable that can take on different values with different probabilities. The circumstances under which a random variable takes any particular value are called an event. In one of these events, one, and only one, particular value of the random value will emerge. One of the analogies commonly used to refer to a probabilistic event is picking a ball with a particular number on it out of an urn full of identical-looking, numbered balls. The balls may all have different numbers, or some of them may have the same number. If they all have different numbers, the probability of picking any particular number is $1/n$, where n is the total number of balls in the urn.

The values that a random variable can take follow what is called a probability density function. The probability density function (call it pdf) shows the probability of occurrence of each possible value of the random variable. Figures 7.1

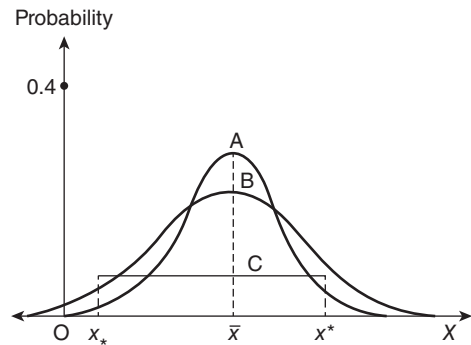


Figure 7.1 Two probability distributions.

and 7.2 illustrate some pdfs. The values the random variable can take are on the horizontal axis, and the probability that each particular value has of appearing in any particular event is read off the vertical axis. The sum of the probabilities of all the possible values is exactly equal to 1.0, which is why we have indicated the vertical scale of the two figures by 0.4 and 0.3 rather than something like 1.0. The total area under each of the curves must equal 1.0, since that area is the sum of the probabilities. The pdfs as drawn refer to continuous random variables rather than discrete ones. The requirement for the area under the pdf to equal 1.0 imposes a relationship between the height of the highest part of the curve and the "thickness of the tails," or the height of the curve at points away from the highest probability. The highest point of the pdf identifies the most likely value of the random variable to occur in any particular event (sometimes called a "draw," as in a draw of a playing card from a deck of cards).³

Pdfs A and B in Figure 7.1 have distinct "central tendencies," with unique, most likely values, and a symmetric falling off of likelihoods of occurrence at values further away from the central tendency. A number of named probability distributions have this characteristic – the normal, the log-normal, and so forth. Pdf C is characterized by uniform probability of each value of x

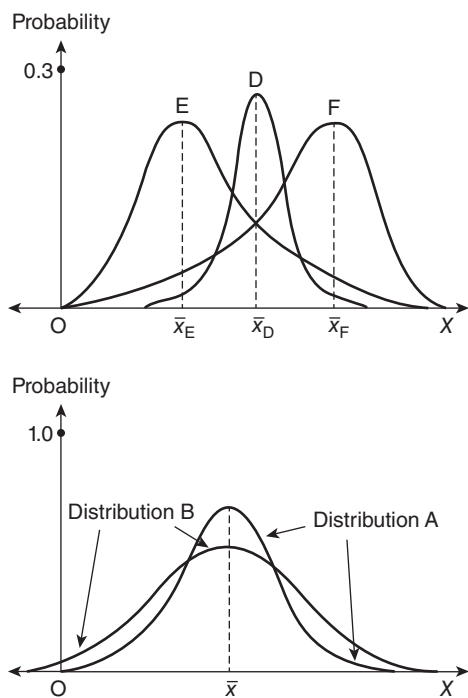


Figure 7.2 (a) Symmetric and skewed probability distributions. (b) A mean-preserving spread.

between x_* and x^* , and zero probabilities for all values less than and greater than those respective values. Distribution A has a tighter distribution of probabilities around its central tendency of $\bar{x}$ than distribution B, which has the same central tendency (which you can call the “mean” of the distribution, the “expected value” of the distribution, or in colloquial terms, the average value of the variable). If we wanted to predict the outcome of an event for both distributions A and B, $\bar{x}$ would be the best estimate we could make without further information, but we would have more confidence in that estimate if we knew that our random variable followed distribution A rather than distribution B.

Figure 7.2(a) shows a very tightly distributed, symmetric distribution in pdf D and two skewed distributions in pdfs E and F. With skewness, either toward the low end of the values (pdf E)

or the high end (pdf F), the most likely values to appear in an event (from a “draw”) are toward one end of the possible values. In pdf E, the mean of the distribution, $\bar{x}_E$, is toward the low end of the possible range of values, while in F the mean, $\bar{x}_F$, is toward the high end of the possibilities. Now is time to introduce a new statistic, the median (the mean, or expected value, is the first “statistic,” or characteristic of a distribution or of a random variable, that we have discussed so far; we will discuss a few other statistics before we’re finished discussing the measurement of risk). The median of a distribution is the value of the random variable, below which (that is, toward lower possible values of the random variable) and above which lie exactly half of the observations in a sample. That is, exactly half of the values of the random variable that occur are smaller than the median value of the random variable, and exactly half are larger. The median and mean coincide in symmetric distributions. In skewed distributions, the mean lies further toward one end of the range of possible values than does the median, which indicates that we’re more likely to see small values than either middling-sized or large values (for positive skewness as in pdf E) or large values rather than middling-size or small values (for the negative skewness of pdf F). Distributions skewed considerably toward very small values of the random variable yield very low probabilities of large values and very high, cumulative probabilities of low values of the random variable. Vice versa for pdfs skewed toward the high values. The Weibull distribution yields such a pattern of probabilities. As you might have surmised, analytical work (that is, pencil-and-paper mathematical manipulations) is a lot easier, and produces “tidier” results with symmetric distributions such as the normal and some log-normals than with asymmetric distributions. Fortunately, a lot of interesting, economic variables are distributed pretty close to normally, particularly when many observations are available on them.

At the risk of sounding trivial, we should discuss the calculation of the mean (calculating the median isn’t as important for our purposes,

and besides, it's more complicated). As every schoolchild knows, she can find the average of an array of numbers by adding them up and dividing by the number of distinct numbers. This looks like $\bar{x} = (x_1 + x_2 + x_3 + \dots + x_{15})/15$, for an array of 15 separate numbers x_i , with subscript i running from 1 to 15. This can be expressed more compactly as $\bar{x} = (1/n \sum_{i=1}^n x_i)$, which is read as "one over n times the sum of x -sub i , where i runs from i equals 1 to i equals n ." This method of calculating an average, or mean, does not use the statistical properties of the random variable. An expression that does is $\bar{x} = \sum_i p_i x_i$, where p_i is the probability of the value x_i coming up in an event, and $\sum_i p_i = 1.0$, which is simply the condition that the sum of the probabilities of all the possible events is equal to one. This expression of the calculation of the mean of a random variable says that that statistic is equal to the probability-weighted sum of the possible values. Using the summation notation, Σ , implies that we are working with the distribution of a discrete random variable rather than a continuous one; the pdfs of Figures 7.1 and 7.2 are for continuous random variables. The probabilistic expression for the mean of a continuous random variable takes us into calculus, so we will content ourselves with showing the arithmetic for the discrete random variable and the picture for the continuous random variable. A final method of expressing the expected value of a random variable is $E(\tilde{X})$, which can be read as the expectation, or the expected value, of the random variable $\tilde{X}$, in which the tilde over the variable is the conventional indication of a random, or stochastic variable.

Now, for the variance, which is our reason for having belabored the calculation of the mean or expected value. The variance is the primary measure of the variability of a random variable about its mean. Pdf D in Figure 7.2 will have the lowest variance of the three symmetric distributions in Figures 7.1 and 7.2; pdf A has the next lowest, followed by distribution B. The arithmetic formulation for the variance is $\sigma^2 = (1/n) \sum_i (x_i - \bar{x})^2$. (For the purists, this is the calculation for a population variance; for a sample variance, $1/n$ is replaced with $1/n-1$, for

reasons we won't discuss.) Using the probabilistic formulation, the calculation is $\sigma^2 = \sum_i p_i (x_i - \bar{x})^2$, which is a probability-weighted average of deviations from the mean of each possible value (or observed value, in a sample). The expectational formulation for the variance is $E(X_i - \bar{X})^2$, or the expectation of the square of the difference between the mean and each possible (or observed) value of the random variable. The variance is sometimes referred to as the "second moment" of the distribution of a random variable; the mean is the first moment; the skewness is the third moment, in expectational notation, $E(X_i - \bar{X})^3$.

The standard deviation, σ , is another commonly used measure of dispersion of a random variable. As the notation itself indicates, it is simply the square root of the variance. As you certainly will have noticed, the mean of a random variable may be either positive or negative, but both the variance and the standard deviation are positive (the square of any number, positive or negative, is positive, and the square root of a positive number is, of course, positive). The measure of skewness, however, may be positive or negative: the third power (in general, odd-numbered powers) of a negative number is negative, while all powers of positive numbers are, of course, positive. Dividing the standard deviation of a random variable by its mean yields a descriptive statistic called the coefficient of variation (abbreviated CV), which may be positive or negative according to the sign of the mean. Sometimes this ratio is multiplied by 100, yielding numbers in the range of, say, 54.3 or 102.8; these values would indicate that the respective standard deviations are 54.3% and 102.8% of the mean. Some writers, myself included, prefer not to multiply by 100. It is a matter of taste, and while some writers do not indicate their calculation, it is usually (but not always) apparent whether they have multiplied by 100.

The measurement of increasing risk has posed a problem, because the simple use of the variance of a random variable has encountered the difficulty that it is easy to construct examples in which a risk-averse individual would prefer a distribution with a greater variance to one

with a smaller variance. The solution to this problem has been the use of what is called a “mean-preserving spread” (of a probability distribution). A mean-preserving spread is a change in the entire probability distribution function such that after the “spread,” the distribution function has more “weight in the tails” of the distribution but the same mean. This is equivalent to any risk-averse individual preferring the previous distribution. Figure 7.2(b) shows an initial distribution, A, and a new one, B, that offers such a mean-preserving spread of the tails; note that the mean (expected value) of both distributions is $\bar{X}$.

7.1.3 Risk and behavior: expected utility

This section develops a model of the relationship between the uncertainties of nature and the uncertainties of consumption. This model leads us into the von Neumann–Morgenstern expected utility model, which we met in Chapter 3. We expand Chapter 3’s discussion of expected utility a bit, to include the concepts of certainty equivalence and risk premiums, and the properties of the individual decision maker’s equilibrium.

Elements of risk

The fundamental issue in risk and uncertainty is that, regarding some particular object of a person’s interest, an array of mutually exclusive things can happen, none of which is perfectly predictable (if one of them were, there would no longer be an array of things that could happen). Additionally, the person himself or herself may take some actions that influence the outcome, but the consequences of any particular action may differ according to which “state of the world” (or “state of nature”) occurs, and since our decision maker can’t know what state of the world is going to appear, neither can she predict the influence of her action. Some structure can be placed over this problem, however, that can help both the decision maker and us as students of decision makers to clarify the situation. We can focus on a set of consequences that depend on *both* the state of the world that occurs and the action that the individual takes before she knows

Table 7.1 Actions depend on events. Adapted from Hirschleifer and Riley 1992, 171, Table 5.1. Reprinted with the permission of Cambridge University Press.

		<i>States of nature</i>		
		$s = 1$	$s = 2$	$s = 3$
<i>Actions</i>	$a = 1$	c_{11}	c_{12}	c_{13}
	$a = 2$	c_{21}	c_{22}	c_{23}
	$a = 3$	c_{31}	c_{23}	c_{33}

which state of the world will emerge. Call this a consequence function, $c = c(a, s)$, where c (the c on the left-hand side of the $=$ sign) is the consequence, a represents the action taken, and s is the state of the world that occurs; the “functional” c on the right-hand side of the $=$ sign stands for the functional relationship between a and s on the one hand (the exogenous variables) and the consequence. Table 7.1 illustrates this.

The columns of Table 7.1, labeled $s = 1$, $s = 2$, and $s = 3$, represent the different states of the world that may occur for some particular decision problem. The rows, labeled $a = 1$, $a = 2$, and $a = 3$, represent actions that the decision maker may take prior to knowing which state of the world will occur. There is no necessity for the number of possible actions, a_i , to equal the number of possible states of the world, s_j . For a farming problem, the three states of the world might be “hot and dry,” “hot and wet,” and “cool and wet.” The actions might be “plant 50% beans, 25% lentils, and 25% wheat”; “plant 25% beans, 25% lentils, and 50% wheat”; and “plant 25% beans, 50% lentils, and 25% wheat.” The nine cell entries are the consequences of the combinations of action i and state of nature j . Thus c_{11} is the consequence for the farmer who plants 50% beans, 25% lentils, and 25% wheat when the weather is hot and dry. We could define the consequence in terms of yield per unit of land, total output, resultant value of the crop harvested, or possibly some other units on which we wanted to focus. In this schema, individuals choose among acts, nature “chooses” among states, and the individuals receive the consequences of the act-state combination.

The states of the world must be clearly observable, and the decision maker must not be able to influence them. In situations in which the decision maker *can* influence the state of the world, we have what is called moral hazard, which we will discuss below. In that case, the uncertainties lie at least in part under the influence of the decision maker.

Next, this decision maker (a farmer in this example) has a set of beliefs about the likelihood of each state of nature occurring. We formulate this set of beliefs as a probability function, $\pi(s)$. We can call any particular probability belief π_s ; $0 < \pi_s < 1$, and $\sum_s \pi_s = 1$. Thus, the value of each probability is strictly between 0 and 1, and the expected probabilities summed over all the possible states (as they are defined for a particular decision problem) must equal 1. If $\pi_s \approx 1$ for one of the states, π_s would be close to zero for the others; this would represent a high degree of confidence, or assurance, on the part of the decision maker about what the state of nature is going to be. This perceived probability distribution would be quite a tight one, like distribution D in Figure 7.2. A considerable degree of doubt about what nature is going to deliver would be represented by $\pi_s = 1/s$ for each of the s states of nature; this would be represented by distribution C in Figure 7.1.

Expected utility

With the definition of the consequences c_{ij} and the probabilities of the state of nature, $\pi(s)$, we can define a preference-scaling function, $v(c)$ that measures the desirability of the different consequences. Notice that the preference scaling function describes alternative consumption bundles, not a bundle with varying combinations of state-of-nature outcomes – only one state of nature occurs. Using the von Neumann–Morgenstern expected utility rule (discussed in Chapter 3), the preference scaling function can give us a preference *ordering* (not a cardinal measure) of acts, $u(a)$, from the individual's preference *scaling* (a cardinal measure) of consequences. To start this connection, we define the “prospect” of an act a . Thinking in terms of a single row of Table 7.1, denote the consequences

of a particular act in each of the different, possible states, c_a , as $(c_{a1}, c_{a2}, \dots, c_{as})$, in which there are s possible consequences, corresponding to the number of states, which we have denoted by s . The state-probabilities expected by the decision maker are $(\pi_1, \pi_2, \dots, \pi_s)$. Then we write the prospect of act a as $a \equiv (c_{a1}, c_{a2}, \dots, c_{as}; \pi_1, \pi_2, \dots, \pi_s)$. (The symbol “ $\equiv$ ” means “is defined to be” rather than “is equal to.”) Using this definition of the prospect of an act a , the von Neumann–Morgenstern expected utility function is $u(a) = \pi_1 v(c_{a1}) + \pi_2 v(c_{a2}) + \dots + \pi_s v(c_{as}) = \sum_s \pi_s v(c_{as})$. The utility of an act is the probability-weighted sum of the preference scalings of the associated consequences. The preference scaling function can be converted into a certainty-equivalent probability associated with the consequence c_{as} , thus making that function interpretable as a probability, and the $u(a)$ function is a cardinal utility function. We have just introduced and used a term we have not defined yet – “certainty equivalence.” We turn to that immediately below.

Certainty equivalence, the risk premium, and risk aversion

The concept of “certainty equivalence” will be a useful one in studying choices under uncertainty. The heart of this issue is the question, “How much hard cash would an individual be willing to exchange for the opportunity (or exposure, depending on one’s viewpoint) to take a well defined (but less than unitary) chance of getting a lot more cash?” Typically the second option is formulated in the literature as a lottery: you pay a certain price for a lottery ticket that has an announced probability of paying off a substantial amount of money (if one wished to consider the matter in terms other than money, that could be done with no change in any of the logic). So, if you are offered a particular lottery – $G(P_h, P_l; \pi_h, \pi_l)$, which means “lottery G has high value P_h with probability π_h and low value P_l with probability π_l ($\pi_l = 1 - \pi_h$)” – what amount of completely certain cash (or in terms of whatever numeraire) would you be willing to accept instead of the chance to get rich quick?

We have already skirted close to this issue in Chapter 3, when we discussed risk aversion and

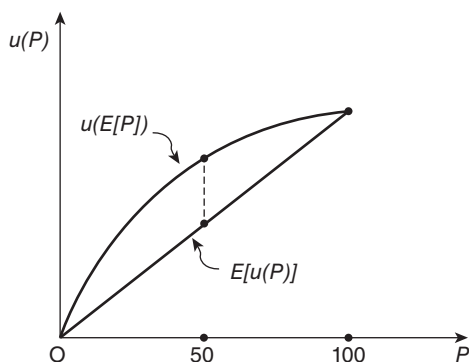


Figure 7.3 Expected utility and certainty equivalence.

found that the expected utility of a particular amount of income need not equal the utility of an equal, expected value of income. Figure 7.3 illustrates the issue again, in terms of a lottery that pays 100 whatever or zero whatever. The straight line connecting $P = 100$ and $P = 0$ (the latter being the origin of the graph) connects all the expected values of this lottery. If the probability of winning the lottery (π_h) is 0.5, we multiply 100 by 0.5 to get an expected value of 50 (indicated on the horizontal axis and marked with a dot on the straight line representing the expected values of the lottery). We have identified the straight line as the expected value of the utility (called “the expected utility”) from the probability-weighted average of winning and not winning the lottery. The dot directly above on the curved line represents the utility this individual would get from a perfectly certain 50 whatever. This difference between the utilities from a sure thing and a risky endeavor (it could be a farming operation or a mercantile endeavor rather than a lottery purchase) sets us up to study more directly the concept of certainty equivalence.

Let’s consider an example in which a person starts our period of analysis with a particular amount of wealth, w_* , and ends the period with w^* , the two being related by the amount of an uncertain gamble, $\tilde{g} : w^* = w_* + \tilde{g}$. We’ll give ourselves a three-part, piecewise-linear utility function that has the risk-averting property of

concavity (the calculations are easier to follow with pencil and paper than for a continuous curve): $u(w^*) = 3w^*$ if $w^* \leq 150$; $u(w^*) = 300 + w^*$ if $150 \leq w^* \leq 225$; $u(w^*) = 525 + \frac{1}{2}w^*$ if $225 \leq w^*$. Utility is a function only of end-of-period wealth, w^* . Now let’s define the lottery or gamble (or farming outcome if one so wishes to characterize it): $G(-50, 30, 100; 0.25, 0.45, 0.30)$. Thus, the decision maker has a 25% chance of losing 50 whatever, a 45% chance of winning 30 whatever, and a 30% chance of winning 100. For a round number, suppose the initial wealth, w_* is 100. Then there is a 25% chance of finishing the period with $w^* = 50$, a 45% chance of finishing with $w^* = 130$, and a 30% chance of finishing with $w^* = 200$. Then expected utility is $E[u(\tilde{w}^*)] = 0.25 \times u(50) + 0.45 \times u(130) + 0.30 \times u(200) = 0.25 \times 150 + 0.45 \times 390 + 0.30 \times 550 = 37.5 + 175.5 + 165 = 378$. So, expected utility is 378. Now, let’s go to the first segment of the utility function, because it includes the value of expected utility of 378 (as long as $u(w^*) \leq 450$, our value will lie along this segment). In this first portion of the utility function, $u(w^*)/3 = w^*$, so we take the quotient $378 \div 3$ to obtain the certainty-equivalent number of whatever of 126 (we’ll call this w_c below). We can see this graphically in Figure 7.4 which graphs the value of the utility function we described earlier in this paragraph against final wealth, w^* .

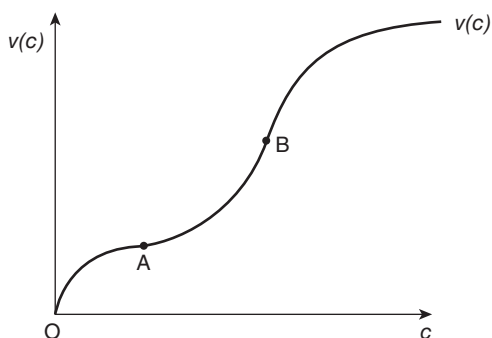


Figure 7.4 Preference scaling function with risk aversion at lower and higher income and wealth levels and willingness to gamble at mid-range.

Suppose that the person we've been considering doesn't have to go out and buy a ticket to this lottery, but already has it – suppose it characterizes a far field on the family farm. What price would it take for her to part with it to a potential buyer? The asking price of this lottery (risky asset) is $p_a = w_c - w_*$, or the certainty-equivalent value minus the initial wealth endowment. How much of this asking price is a risk premium? Define the risk premium as $E[\tilde{g}] - p_a$, or the expected value of the gamble minus the asking price of the lottery (asset). We can calculate $E[\tilde{g}]$ as $-50 \times 0.25 + 30 \times 0.45 + 100 \times 0.30 = -12.5 + 13.5 + 30 = 31.5$. Then the risk premium, p^* is $31.5 - 26 = 5.5$. The asking price, p_a , can be positive or negative (you might be willing to pay somebody to take some types of property off your hands), but the risk premium, p^* , will always be positive.

A decision maker's possession of the characteristic we have called risk aversion does not imply that such a person is desperate to avoid all risks. If people are risk averse, they do not value risk for itself, so they must be compensated by some premium over the simple, expected return of an uncertain venture. An alternative insight into the thinking behind risk aversion is that a person with one of these concave (that is, risk averse) utility functions need not have any particular dislike for uncertainty per se, but instead might feel that he or she can do better in the long haul by playing it safe presently, increasing the likelihood of preserving resources to take advantage of more favorable opportunities in the future.⁴ It is suspected that the shape of the utility functions (as functions of income or wealth) may be somewhat more complicated than the simple form that yields a continuous increase in wealth, but at a declining rate, such as Figure 7.3 uses. Figure 7.4 illustrates a shape that would account for strong risk aversion at particularly low levels of income and wealth and at upper middle and high income and wealth, but a willingness – indeed a preference – to gamble in the upper end of low incomes and lower end of middle incomes.

Individual equilibrium under uncertainty

You may think back to our analysis in Chapter 3 of the optimum conditions for which a consumer

searched in the absence of uncertainty and recall that that involved equalizing ratios of goods prices and marginal utilities from the consumption of those goods. We will find a similar structure of relationships in the expected utility case, but with some important differences. Expected utility, $u(a)$, is defined ordinally over actions that lead to prospects of consumption, so let's recall the definition of the prospect: $a \equiv (c; \pi) = (c_1, c_2, \dots, c_s; \pi_1, \pi_2, \dots, \pi_s)$, in which the c_i are the state consequences and the π_i are the state probabilities; the a is omitted from the subscript of the consequences. Recall that the consumer doesn't get to choose the amounts of, say, c_2 and c_5 that she gets to consume together, as she would with, say, lentils and olives. Consumption c_2 is everything she gets to consume if state 2 occurs, and c_5 is everything she gets to consume if state 5 occurs. When state 2 occurs, she gets c_2 , and c_5 is nonexistent. Consequently the *ex ante* choices the decision maker faces are the possible consumption levels and combinations she is willing to accept across the different states, only one of which will occur in each period.

We will examine a simplified case in which there are only two possible states and the consumption bundle is composed of a generalized substance we will call "food." The utility function is $u = \pi_1 v(c_1) + \pi_2 v(c_2)$. We study the properties of this formula graphically in Figures 7.5 through 7.7.⁵ The individual begins with the endowment of entitlements to consume $\bar{c}_1$ if state 1 occurs and $\bar{c}_2$ if state 2 occurs, identified by the point $\bar{c}$ in the southeast corner of Figure 7.5. A 45° line through the origin of the graph shows all the combinations of c_1 and c_2 that are exactly equal to each other, the amounts that would be chosen in the absence of any state uncertainty. Hence the line is called the certainty line. Next, we draw a line, LL , through the endowment point $\bar{c}$ that has a slope equal to the negative of the ratio of the two state probabilities, $-\pi_1/\pi_2$; the slope of this line also happens to equal the ratio $\Delta c_2/\Delta c_1$. Consequently, $-\pi_1/\pi_2 = \Delta c_2/\Delta c_1$, or $-\pi_1 c_1 = \pi_2 c_2$, which says that the probability-weighted consumption levels in the two states are the same everywhere along this line. The slope of any indifference curve between c_1 and c_2 (which has a slope of $-\Delta c_2/\Delta c_1$), when it crosses the certainty line, is equal to π_1/π_2 ; consequently, we have the

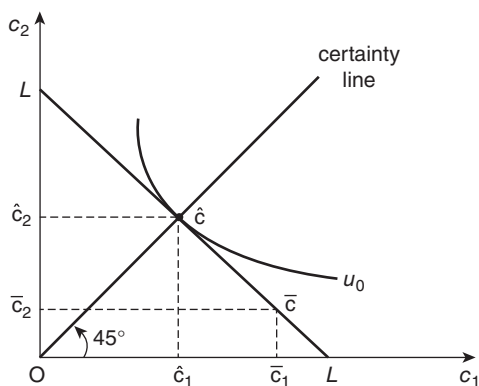


Figure 7.5 States of the world: certainty equivalence and a consumer's willingness to substitute between states.

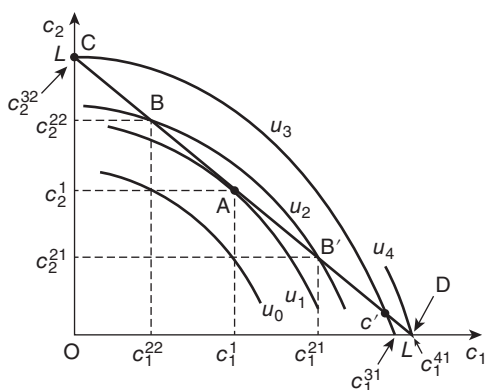


Figure 7.6 Choosing risky consumption.

indifference curve u_0 tangent to LL at the intersection of LL and the certainty line, denoted by the certainty consumption point $\hat{c}$ in Figure 7.5, which gives state consumption levels of $\hat{c}_1$ and $\hat{c}_2$. The marginal rate of substitution (MRS) between c_1 and c_2 at any point along the indifference curve is $[\pi_1(\Delta v / \Delta c_1)] / [\pi_2(\Delta v / \Delta c_2)] = -\Delta c_2 / \Delta c_1$.

Let's digress a moment into the subject of risk aversion. The shape of indifference curve u_0 in Figure 7.5 implies risk aversion, which itself implies diversification. Thinking back to Figures 7.3 and 7.4, the curvature of the preference scaling function $v(c)$ (the risk averse segments of $v(c)$ in Figure 7.4) has the property that $\Delta(\Delta v / \Delta c) / \Delta c < 0$: that is, decreasing – but still positive – marginal increments of utility as

c is increased.⁶ A positive sign of this curvature characterizes the stretch of $v(c)$ between points A and B in Figure 7.4, over which consumption or income range we characterized the consumer's attitude toward risk as risk-prefering. This curvature of $v(c)$ translates into the convex shape of the indifference curves in prospects, $u(a)$: as more, say, c_1 , is substituted for c_2 , the increments of c_1 that have to be substituted for c_2 have to be larger as we approach larger ratios of c_1 to c_2 . To see that diversification between c_1 and c_2 is implied by this shape of the indifference curve u_0 , and hence by the curvature $v'' < 0$ in the preference scaling function, we show a concave indifference curve u_1 reflecting a preference for risk, and a shape of a preference scaling function characterized by $v'' > 0$ in Figure 7.6. Our indifference curves u_i are concave to the origin (bowed out). Utility level u_1 reaches a tangency with line LL (which we treat as a budget line for current purposes). The budget represented by LL will allow consumption of the pair c_1^1 under state 1 and c_2^1 under state 2 (labeled point A), but the consumer could afford to reach a higher level of utility by shifting her consumption prospects to either the combination c_1^{22} and c_2^{22} (point B) or c_1^{21} and c_2^{21} (point B'), both of which contain a higher ratio of one or the other of the two prospects. However, this consumer could reach a higher level of utility and still stay within the budget constraint by consuming no c_1 at all and c_2^{32} of c_2 (point C, on the c_2 axis), which implies death if state 1 occurs – quite a gamble according to most people's ways of thinking. She can do better by herself yet, however, if she chooses to thrive if state 1 occurs and die if state 2 occurs by reaching utility level u_4 and consuming c_1^{41} of c_1 and no c_2 (point D, on the c_1 axis). Points C and D are called “corner solutions” in the technical lexicon. Sometimes corner solutions are interesting, but in this particular case, these corner solutions simply highlight the unlikelihood of completely general risk-preference, or $v'' > 0$ throughout the entire range of a preference scaling function instead of only over a relatively small stretch of levels of income or consumption c as shown in Figure 7.4.

Returning to our main story line, we continue the exposition of the risk-averse individual's optimal choice of prospects in Figure 7.7, in which we omit indifference curve u_0 but mark

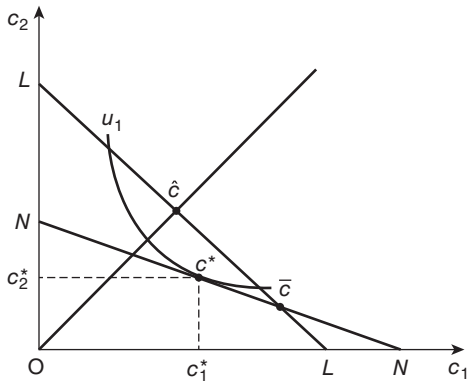


Figure 7.7 State-dependent consumption.

the tangency point on the probability ratio line LL as $\hat{c}$ and the endowment point on LL as $\bar{c}$. We add a budget line NN , which depicts the relative price p_1/p_2 , the relative prices of the generalized “food” prospects under the alternative states. At the endowment point $\bar{c}$, $p_1/p_2 = \pi_1/\pi_2$. Line NN shows all the combinations of c_1 and c_2 our decision maker could afford by exchanging some of her initial endowment. At the price ratio shown and with the preferences embodied in the shape of the indifference curve family u_i , she would like to give up some consumption in the event of state 1 to live a little better should state 2 occur. (Only a risk-neutral consumer would want to consume at the endowment point $\bar{c}$, where the ratio of prices equals the ratio of probabilities.) Indifference curve u_1 is parallel to the unshown curve u_0 and is tangent to the budget line NN at c^* , with the combination of consumption prospects c_1^*, c_2^* . Note that, relative to the endowment point, c^* is closer to the certainty line, indicating that the consumer has chosen her consumption prospects in a way that reduces her risk, but in equilibrium, she still accepts some risk, indicated by the fact that c^* is off the certainty line. The tangency of indifference curve u_1 with budget line NN has the property that $\pi_1 v'(c_1)/\pi_2 v'(c_2) = p_1/p_2$.

At this point in the discussion, we’re going to give another name to c_1 and c_2 and their respective prices, p_1 and p_2 . State-dependent consumption bundles c_i are called “contingent claims” or “state claims”; a person who possesses the right to consume bundle c_1 should

state 1 occur has a claim on consumption in that state of the world, but the claim (or right) is contingent on the particular state of the world occurring. Thus the term “contingent” or “state claim” for the c_i . Correspondingly, the prices p_i are called “contingent-claim prices” or “state-claim prices.” For a utility maximum, the ratio of probability-weighted marginal utilities (the $v'(c_i)$ using the more compact notation for Δ -differences) to the ratio of their state-claim prices. Rearranging this relationship, we find that, in general, the ratios of probability-weighted marginal utilities of state claims to their prices should be equalized for all state claims: $\pi_1 v'(c_1)/p_1 = \pi_2 v'(c_2)/p_2 = \dots = \pi_s v'(c_s)/p_s$ for all s possible states. This can be rewritten as a set of ratios of expected marginal utilities to ratios of state-claim prices: $\pi_i v'(c_i)/\pi_j v'(c_j) = p_i/p_j$, for all states i and j . This relationship is known as the Fundamental Theorem of Risk-Bearing. Notice that the ratio of state claim prices p_i/p_s equals the ratio of state probabilities π_i/π_s only for the case where $v'(c_i) = v'(c_s)$, known as the “fair gamble”; in this case, relative price (budget) line NN and the probability ratio line LL would coincide.

Before we leave expected utility and contingent claims for a while, let’s look at the effects of changing probability ratios, state-claim prices, and risk aversion on equilibrium movements from an initial endowment to a utility-maximizing state-claim combination. In Figure 7.8, we have increased the slope of line LL to $L'L$, raising

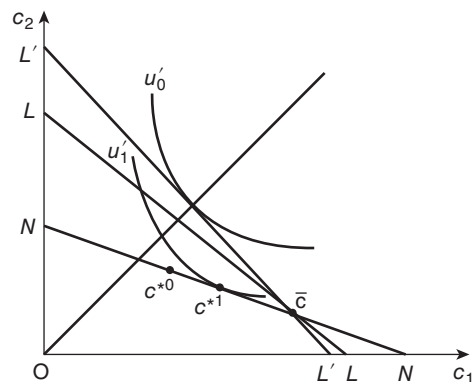


Figure 7.8 Effect of changing probabilities of the risky states.

the ratio π_1/π_2 : state 1 is now more likely to occur. We begin from the same endowment $\bar{c}$ that we had in Figures 7.5 and 7.7. We reproduce the optimum state-claim combination from Figure 7.7 as c^{*0} . The new family of indifference curves cannot be parallel to the original family; we denote the new family as u'_i . The new, utility maximizing combination of state claims, c^{*1} , lies to the southeast of c^{*0} along budget line NN . State 1 being more likely to occur, we are willing to get by with the prospect of less should state 2 occur: the substitution of c_2 for c_1 is more expensive in the sense that the c_2 being substituted for c_1 is less likely to occur. Consequently, we substitute less c_2 for c_1 .

In Figure 7.9, we increase the relative price of state-1 consumption, p_1 , twisting the line NN around the endowment point $\bar{c}$ to $N'N'$. The family of indifference curves in Figure 7.9 is the same as that in Figures 7.5 and 7.7, and we show only curve $u^{1A} > u_1$ of Figure 7.7. The new equilibrium combination of state claims, c^{*1} , is composed of a combination of less c_1 and more c_2 . The rationale is quite simple: c_1 costs more now relative to c_2 . Figure 7.10 changes the preferences of the consumer to less risk aversion: the new indifference curve u_0^* , tangent to LL at the intersection with the certainty line, has less curvature than the old indifference curve, u_0 . At any c_1 coordinate, consumers need a larger increment of c_2 to be compensated for a unit reduction in c_1 (draw an imaginary, vertical line

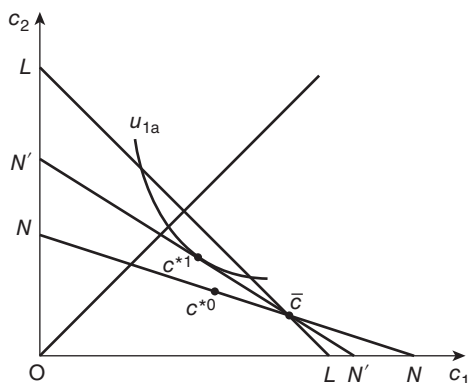


Figure 7.9 A higher price of state-1 consumption.

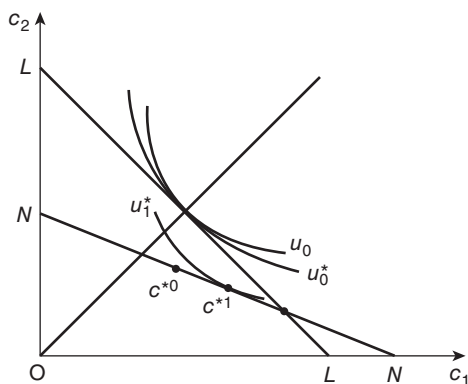


Figure 7.10 Lower risk aversion on the part of a consumer.

through u_0 and u_0^* and compare the slopes of the two indifference curves at their intersections with that line); equivalently, they require a smaller increment of c_1 to compensate for a unit reduction in c_2 . Expressed alternatively, the slope of the new indifference curve at all points away from the tangency with the π_1/π_2 line is closer to the slope of that relative probability line than is the original, more risk-averse indifference curve. Consumers with this new preference system are closer to being willing to accept "fair gambles"; they require less of a risk premium to accept any risky venture.

7.1.4 Risk versus uncertainty: the substance of probabilities

It is common to observe a distinction between risk and uncertainty such that risks involve well known probabilities (for example, it's early summer in Chicago, the sky is cloudy at midday, therefore the likelihood of rain is $x\%$ – in my opinion) while uncertainties involve quantities about which actuarial probabilities are unknown for want of observations or actually are unknowable (for example, will Persia and the Athenians go to war next year? Will there be a silver strike near Tarsus next year?). This distinction parallels a long-running debate within the discipline of statistics regarding the objectivity or subjectivity of probabilities: under what circumstances can

we distinguish between objective and subjective assessments of probabilities? The subjectivists generally have prevailed in that debate, and certainly within the social science applications of statistics and probability theory, the ultimate subjectivity of probability assessments has carried the day. People learn about the probability of a particular coin coming up heads and tails, or the chances of any particular face coming up on a die, by observing tosses with that coin or casts of the die. They may begin with a prior assessment that most coins will have about a fifty-fifty chance of coming up heads or tails and most dice have about a one-in-six chance of each face appearing – provided the coin is “fair” or the die is not “loaded.” The only means of determining the honesty of the coin or the die is to observe its performance in repeated trials. If a coin comes up heads 812 times out of 1000, an observer probably would revise his prior assessment that the coin is fair and adjust the probabilities he assigns to heads and tails accordingly. The probability assessment that a colonizing farmer assigns to the possibility that he will get the same yield per acre in wheat that he got at home will evolve as he learns about the new environment.

An individual's assessment of current truths and future likelihoods is conditioned by his imperfect information about the present and by his experience and understanding of the past. These two sets of knowledge form the filter through which he or she passes new, “raw data” to generate new subjective judgments about probabilities of various events coming to pass in the near or far future. Different agents can draw different conclusions from the same data because of differences in either or both current information and past experiences – essentially, what they “know” (believe) at the time the new information arrives.

Aspects of the future – or regarding another place or person – about which we do not know “for certain” have probabilities attached to them. We decide on the probabilities. Many contingencies we may not have given much thought to, so if asked we would not immediately have a prior estimate. If we care so little about some particular contingency that we are unwilling to think

seriously to make an estimate, the default probabilities with which we are left are $1/n$ for each of the n possible states the contingent event could take. Of some things we are quite certain. This could mean either that we assign a probability of 1.0 to one of the possible outcomes of that event, or that we assign a probability arbitrarily close to 1.0 because we haven't really thought about the possibility that our belief isn't true. One view of probabilities is that each and every probability is conditional on some prior information; this view leads to the conditional probability, which is equivalent to a belief that, with the information currently at one's disposal, the probability of state s occurring for some contingent event is “such and such.” With some additional information, we may change our assessment of the probability either upward (to become more likely) or downward (to become less likely). Or, new information that arrives we may consider not particularly informative and not sufficient basis for changing our previous judgments very much at all.

This approach to probability, and to knowledge and beliefs in general, should prove quite compatible with the study of ancient societies, for whose beliefs and understandings the scholar must be careful not to substitute any more of his own than he or she can avoid. Ancient religions, technologies, and scientific understandings of both nature and society all changed – to avoid the loaded term “evolved” – over the centuries. This model of updating prior beliefs with new information, which itself is evaluated according to its conformance with expectations and previous beliefs, is flexible enough to accommodate a wide range of beliefs and levels of technology. Without presuming that ancient decision makers, as either consumers or producers, were skilled statisticians, we can apply contemporary probability models to sharpen our understanding of how they might have behaved. The same framework applies to our own evolving understanding of their understandings: prior beliefs, new information, filtering, new understandings.

Having just introduced the outlines of Bayesian probability theory, which is the dominant outlook on uncertainty in contemporary economics, we

will turn to some additional formality regarding knowledge and information in the following section.

7.2 Information and Learning

I begin this section with a bit of economic terminology about information and an informal discussion of some ways in which economic agents deal with information. The second subsection develops somewhat further the logic of how new information is used to revise subjective probabilities regarding states of the world. The final subsection discusses problems that can arise in two settings: soliciting information from experts (*à la* the Delphic Oracle) and the use of information in groups, where beliefs may differ and conflicts of interest may exist.

Some of the notation in this section appears daunting, even when it is serving essentially simple concepts. Frequently we need to keep track of at least two alternative states of the world, two individuals, and two “things” that can be exchanged. There are consumption items, contracts that confer rights to consumption, and probabilities. Consequently the subscripting and superscripting occasionally looks like it is running riot. I ask the reader’s patience. I explain the notation, but using it does permit us to shorten sentences by as much as a half-page occasionally, and working back and forth between the notation in the occasional formulaic representation and its verbal interpretation may actually improve the understanding of the expositions.

7.2.1 *The structure of information*

There are, undoubtedly, many ways to think about information and knowledge. We offer a characterization here that emphasizes distinctions between stocks and flows, scarcity and nonscarcity, intentional and unintentional. These aspects of information can be combined to yield a useful perspective on individual and group behavior *vis-à-vis* information.

At any time, a stock of accumulated information is available in a society. This stock we will call

knowledge. Individuals generally possess some but not all of the extant knowledge at any one time. Access to the extant stock of knowledge may be costly for one of two principal reasons: retrieval of the knowledge, and transmission from one agent to another. It may require effort to determine where some knowledge is stored, and to bring it out of storage to the site or agent of its demand. Some knowledge may not be immediately transparent and may require extensive interaction between agents or lengthy study by an agent working alone; foreign languages are an obvious example.

Increments to the stock of knowledge, either to society as a whole or to individuals possessing their individual stocks of knowledge, can be said to arrive as messages or news. We can think of messages as intentional communications and news as more general revelations. Messages are the products of “message services,” which can be purchased or hired. Messages themselves cannot be bought, only message services. For example, you could subscribe to a weather forecasting service and receive the news every day about what the day’s weather was forecast to be, but regardless how much you paid, you could not buy the message that tomorrow’s weather is going to be warm and sunny; it’ll be whatever it’ll be. When you subscribe to a message service, you get whatever messages come out of the service. Message services can vary in the quality of the messages they provide, quality consisting of accuracy, detail, and timeliness. Some services may be more skilled and consequently probably more costly, while others may be cheaper because they apply fewer resources to message acquisition.

Some intended communications are unintentionally inaccurate, some intentionally so, particularly in the cases of messages sent strategically between individuals or between organizations. These intentional releases of information between interacting parties are commonly called signals; they may be deliberately or incidentally incomplete; they may be deliberately deceptive.

Some information may be private, some public. The difference refers to the public/private good distinction in that some information may be retained by its holders with a minimum of

leakage to others, while other information is harder to keep secret. Sometimes attempting to use one's knowledge will reveal it, totally or partially. This revelation may occur through observations of the transactions of informed persons. By such means, uninformed persons may learn at least part of what informed persons may have paid to learn. This potential for "leakage" of knowledge acquired through costly means can pose a serious limitation on the incentive to devote resources to knowledge acquisition. People may possess private information about themselves that in some cases they would like to convey to others – for example, that they are competent and diligent – and in other cases that they would like to keep to themselves – that they may prefer to shirk than work hard, that they are wealthy when the tax collector comes around.

The information relevant to a particular transaction may be held asymmetrically between or among the partners to the transaction. Buyers and sellers, whether in a simple transaction of a consumption or durable good or in a complex contract, may not have the same information available to them. At least some of them may prefer to keep things that way. We will discuss asymmetric information more extensively in section 7.4.

The production of information is not a particularly simple economic activity. First, it suffers from a public good problem. If one person produces some information and tries to sell it to another, the potential buyer is likely to want proof that it is valuable to her. To offer the required proof, the producer / seller may have to reveal enough of the information that the seller no longer needs to buy it to effectively possess and use it. If such failures of excludability are complete enough, this characteristic can seriously retard the costly production of information. Patent and copyright protection offered by contemporary governments offer limited protection to investments in information production, but both solutions carry the undesirable property that they may restrict use of the information to less than the optimal amount. Second, information also suffers from what is called "the problem of the commons." At any

point in time, given a society's knowledge base, there is a stock of potential inferences that can be extracted, and this stock is available to the entire society – or to anyone in the society with sufficient resources to "mine" the stock. As no one owns the potential inferences, there will be a tendency for diverse agents to expend more than the optimal volume of resources in the effort of inference production – "knowledge mining," as it were. This is analogous to fishermen overfishing when no agent is able to make effective a claim of ownership of the fish. This would not be a problem in a privately owned and stocked pond from which fishermen could pay to fish, but it is a problem in open seas and other unrestricted waters. In such situations, the inputs used in the production effort – either fishing or informational inference generation – must be paid the value of their marginal cost but they are used to the point where they are yielding only their average product, which is below their marginal product. The equalization of marginal cost and average product results in excessive use of a factor.

As we discussed in section 7.1.4, at any particular time, people hold beliefs that are subjective inferences from their knowledge and that form the basis of their actions. As they receive new information in the form of either specialized messages or incidental signals, they may revise their beliefs and, accordingly, their actions. This process can be formulated as a tripartite structure involving the holding of a set of beliefs about probabilities of states of the world, the entry of new information, and the interaction of the previous beliefs and the new information to produce new beliefs. In the lexicon of Bayesian probability theory, these are prior probabilities, a likelihood function for evaluating new information, and resultant posterior probabilities. We turn now to that subject.

7.2.2 *Learning as Bayesian updating*⁷

We have discussed at several previous points how learning can be characterized as a process of using new information to revise previously held, subjective probabilities of various states of the world occurring. To support the exposition, we

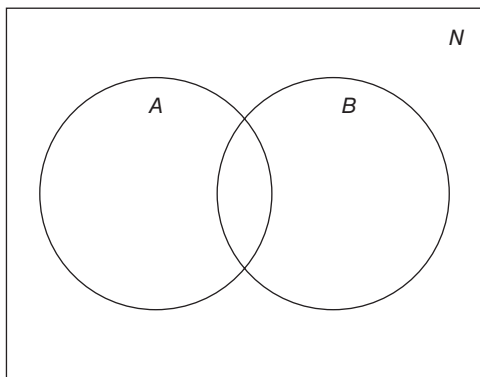


Figure 7.11 Overlapping events.

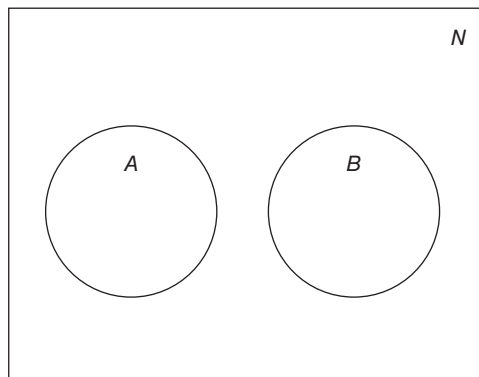


Figure 7.12 Independent events.

develop three probability concepts: joint probability, conditional probability, and unconditional probability. The box N in Figure 7.11 represents all the possible outcomes (states of the world that we have identified as possible in some particular problem, say, “What will the weather be like next week?”) The circles A and B are two distinct events; say A is the event “hot weather” and B is the event “rainy weather.” Everything else outside A and B is not hot and not rainy (cold and dry?). The area of overlap of A and B is both “hot and rainy.” The unconditional probability of hot weather (the notation is $P(A)$) is the area A divided by the area N , and the unconditional probability of rainy weather, $P(B)$, is the area B divided by the area N . The joint probability of hot *and* rainy weather – the notation is $P(AB)$ – is the area of the overlap of circles A and B , divided by N , or the product of $P(A)$ and $P(B)$.⁸ Events A and B in Figure 7.12 do not overlap and hence are mutually exclusive; they therefore cannot be independent.

A conditional probability is the probability of one event, given the information that another event has occurred. The symbol for a conditional probability is $P(A|B)$: “the probability of A , given B .” This conditional probability is $P(A|B) = P(AB)/P(B)$, or the joint probability of A and B , divided by the unconditional probability of B . Dividing by the probability of B rescales the joint probability of A and B from the entire set of events to just the portion of them represented

by the probability of B , which of course yields a conditional probability greater than the joint probability. It will not generally be the case that $P(A|B) = P(B|A)$. There is no temporal implication in a joint probability; it is not necessary for B to have already happened to make the probability of A conditional on B . Events A and B could occur simultaneously, or A could happen before B just as well as B before A .

Now, we have the equipment to introduce the model of revising initially held probabilities with new information. We start with a matrix of messages in different states. As a vehicle for this exposition, we develop the example of reports home to Assur from family trading agents in Boğhazköy. I have not actually taken these messages from the recovered tablets, but rather use that situation as a story around which to build the logic of this model. Suppose the head of the family trading firm in Assur wants to assess how the Hittite market is going to unfold over the next year so he can decide whether to increase his fixed investment at the site and his productive capacity at home. No one, of course, actually *knows* how that market is going to develop, so the boss will use routine information sent back home as his messages. Table 7.2 sets up in matrix or tabular form an array of possible messages and possible states of the world (or states of the market in Hatti).

The entries in the cells of Table 7.2 are probabilities – joint probabilities of being in

Table 7.2 Joint probability matrix. Adapted from Hirshleifer and Riley 1992, 171, Table 5.2. Reprinted with the permission of Cambridge University Press.

	Messages					Probabilities for states $\sum_m j_{sm} = \pi_s$
	<u>1</u> sales up, orders up	<u>2</u> sales up, orders steady	<u>3</u> sales steady, orders down	<u>4</u> sales down, orders up	<u>5</u> sales down, orders steady	
States <u>1</u> market expanding next year	j_{11}	j_{12}	j_{13}	j_{14}	j_{15}	π_1
<u>2</u> market steady next year	j_{21}	j_{22}	j_{23}	j_{24}	j_{25}	π_2
<u>3</u> market stagnant next year	j_{31}	j_{32}	j_{33}	j_{34}	j_{35}	π_3
<u>4</u> market declining next year	j_{41}	j_{42}	j_{43}	j_{44}	j_{45}	π_4
probabilities for messages $\sum_s j_{sm} = q_m$	q_1	q_2	q_3	q_4	q_5	$\sum_s \sum_m j_{sm} = 1.0$

state s and receiving message m . For instance, entry j_{11} is the probability the boss assigns to receiving the message that sales and orders are both up in a state of the world in which the Hatti market is expanding. By all accounts, such a probability should be relatively high. He might hold a similarly valued probability for j_{21} (the Hatti market steady next year), but his anticipations of receiving this message in states 3 and 4 (stagnant and declining) would be relatively low. The sum of all the j_{1m} probabilities, where m runs from 1 to 5 for the five possible messages, equals the probability assigned to the occurrence of state 1, π_1 . Similarly, if we add all the j_{s1} probabilities of receiving message 1, we obtain the total probability of receiving message 1. Both the sum of the state probabilities and the sum of the message probabilities equals 1: one of the states *must* occur and one of the messages *must* occur. The π_s are the prior probabilities regarding the occurrence of the states, and the q_m are the prior probabilities assigned to receiving the messages.

These are the beliefs prior to the receipt of any new information.

Two conditional probability matrices derive from this joint probability matrix. First, we can divide each j_{sm} entry in each row by the corresponding state probabilities π_s for that row, to obtain the conditional probability of any message conditional on the state. In this new matrix, the probabilities in each row sum to one. This is the likelihood matrix, which represents the message service discussed in section 7.2.1. The probability entries in the cells of the likelihood matrix are denoted q_{m-s} , indicating the probability of the alternative messages, given each state. This may be an objectively derived set of probabilities – that is, derived from observed frequency data – but it need not be – it may, in fact, be “reasonable beliefs” based on nonquantitative information.

The other matrix contains the conditional probabilities of each state, given a particular message. We get it by dividing the j_{sm} in each column by

the corresponding message probability q_m . The sum of the j_{sm} down each column of this matrix equals one. The entries down a column in this matrix are the probabilities a person ought to assign to states of the world after having received the message associated with that column. The cell entries in this matrix are denoted $\pi_{s,m}$, for the probability of state s , given message m . This matrix of probabilities shows all the possible posterior probability distributions that the boss could come up with; the one he picks depends on which message he receives. One of the columns of this matrix will represent his new set of beliefs that motivates his subsequent actions.

The revision process works as follows. The boss starts out with two sets of prior probability distributions: π_s for states and q_m for messages. Second, he sets up his likelihood matrix (or likelihood function), in which he sets up the message probabilities to be inferred for each state. He multiplies the “state priors,” π_s , by the corresponding elements of the likelihood function to get his joint probability matrix. The column sums are his message probabilities, with which he can compute his “state posteriors.” Third, he obtains a new message – he looks to the appropriate message column in his likelihood function. He looks at the column of the posterior probability matrix associated with the particular message he received and bases any action he takes on the probability distribution associated with the particular message he received. The probability revision process can be summarized by the expression $\pi_{s,m} \equiv j_{sm}/q_m \equiv \pi_s q_{m,s}/q_m$, which is one expression of Bayes’ theorem.

Figures 7.13 through 7.16 illustrate this procedure. Figure 7.13 shows the steps involved in updating previously held beliefs via the evaluation of new information. In the top line of Figure 7.13, initially held beliefs are simply the result of a previous updating procedure and are contingent on information available at the time. All probabilities are conditional on some information – the unconditional probabilities $P(A)$ we denoted from Figure 7.11 are in actuality conditional on N (as in $P(A|N)$), which must always occur in one form or another, so there is a division by 1.0 in the conditional probability formula, which is

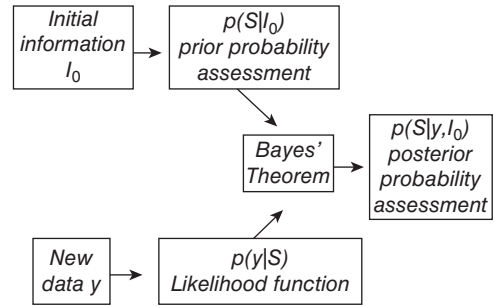


Figure 7.13 The Bayesian updating process. Adapted from Zellner 1971, 10, Figure 1.1. This material is reproduced with permission of John Wiley & Sons, Inc.

commonly omitted. In the parallel, bottom line of steps, new data are collected and evaluated with the likelihood function. The new data are faced with the question, “Given our assessment of the probability distribution of states S , what is the probability that we would have gotten the observations y ?” The less likely were the observations y to be observed, the more informative are the new data; if they were exactly what would have been expected given the prior state probabilities, there is no reason to revise those priors. Anyway, this new set of probabilities is filtered through Bayes’ theorem to yield a revised set of posteriors, which may or may not differ much from the priors.

Figure 7.14 shows the operation of the likelihood function. The rightmost probability distribution is the prior probability distribution – prior to the arrival of the new data and the filtering of it through the likelihood function. The likelihood function is the dashed distribution curve⁹ and is “humped” strongly to one end of the states, indicating that it contains strongly different probabilities than those initially held. The center of the probability mass is higher than that associated with the prior probabilities, indicating greater confidence in the central estimate. The multiplication of the likelihood function by the vector of prior state probabilities yields the intermediate posterior distribution, which has a tighter probability mass than the priors but looser than the likelihood function, which,

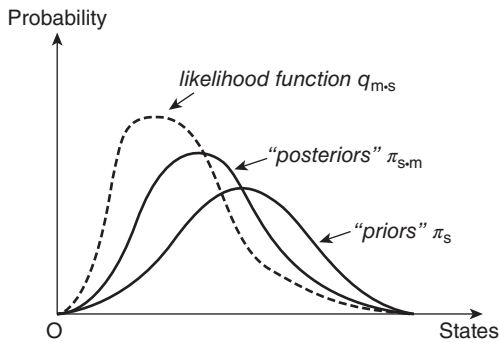


Figure 7.14 Subjective probability distributions before and after new data arrive. Adapted from Hirschleifer and Riley 1992, 176, Figure 5.1b. © 1992 Cambridge University Press.

after all, is only a single “test”; repeated tests yielding the same likelihood function (“pretty much” or “more or less”) would pull the original prior distribution much closer to the shape of the likelihood function’s distribution. The most likely state (or the belief about which state is most likely) is pulled to the left in this case, toward the most likely state estimated by the new data in the likelihood function. (We have implicitly ordered the five states in our example symmetrically around the one believed most likely to occur.) Figures 7.15 and 7.16 show examples of totally uninformative new data, which do not change the priors at all, and highly informative new data that essentially reinforce the original priors,

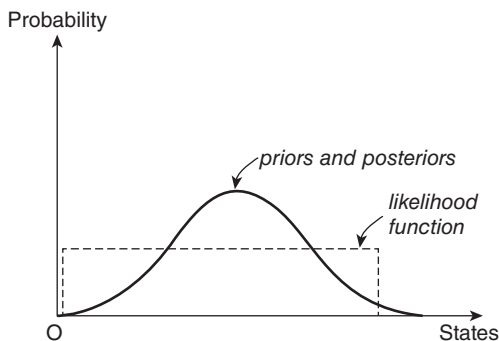


Figure 7.15 Uninformative new data.

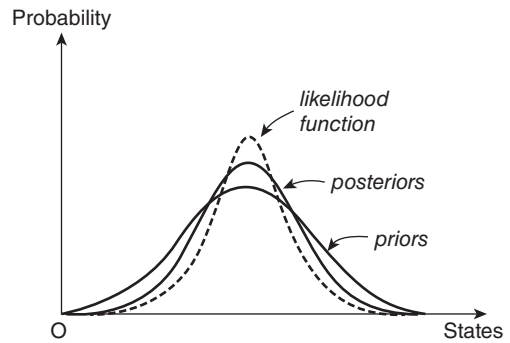


Figure 7.16 Highly informative new data.

tightening the variance around the perceived most likely state.

In this revision process, an agent does not abandon his or her original beliefs, but rather gradually modifies them. A long-held set of beliefs which had been reinforced by repeated observations over a period of years would yield a tight distribution of prior probabilities, rather like the likelihood function of Figure 7.16. Repeated “experiments” would reduce the variance of the “estimates” of probabilities. Consequently, a single, equally tight likelihood function would not “move” such a prior pdf all that much. What are called “diffuse priors,” distributions that look like the likelihood function of Figure 7.15, would be relatively easy to move with new information coming through likelihood functions. This accounts for the logic behind what can be called change-resistant beliefs.

The Bayesian approach to revising probabilities also yields an alternative to the actual *testing* of hypotheses in the form of either rejecting or accepting them. We could assign probabilities to hypotheses – for example, the hypothesis that the market in Hatti will improve next year and the alternative hypothesis that it will decline. The boss of our Assyrian family trading firm might begin by assigning equal probabilities to the two hypotheses (probabilities in the sense of “reasonable beliefs”). Following the appearance and examination of some new information, he may revise his priors of $1/2$ – $1/2$ to $3/4$ for “the market will improve” and $1/4$ for “the market

will decline.” The posterior odds in favor of the former hypothesis have changed from 1-to-1 to 3-to-1, and the boss could be justified in saying that the former hypothesis is “probably true,” without feeling the need to actually test the hypotheses and reject one of them. This Bayesian process is one of comparing hypotheses rather than testing them.

7.2.3 *Experts and groups*

Having just introduced the concepts of the likelihood function and the use of it to update beliefs, we will give that concept some further immediate exercise with applications to two problems, one involving the solicitation (and possibly, but not necessarily, the production) of information, the other the use of information by multiple agents attempting to make a single decision. In the information production / solicitation problem we will consider the use of an expert to provide information, a situation equally applicable to the contemporary use of a business or political consultant and ancient appeals to the Delphic Oracle prior to Greek civic endeavors and similar consultations with priests and seers in Egypt and the Mesopotamian polities (reading livers and other entrails, and so forth). In the multiple-agent information problem, the decisions of groups – business or political organizations – frequently hinge on the opinions of more than one person. The individuals with the capacity to influence decisions may bring different prior beliefs, different likelihood functions, or different interests into the discussion preceding the decision. All these differences may influence how – or even whether – the decision is made.

Contracting with experts for information

We begin with the expert first. Soliciting information from an expert is tantamount to asking that person to use his or her likelihood function to filter observations and report the results to the person or group “hiring” the advice. The client faces two possible problems in dealing with the expert: can the expert be induced to tell the truth, and can she be induced to take the question seriously enough to be accurate?

Consider truthfulness first. The expert will understand the Bayesian structure of the information problem: she will be well aware that the client asking for the opinion will have a set of prior beliefs – possibly “diffuse” beliefs, which are quite close to “knowing nothing” – and that whatever opinions she delivers will be filtered through the client’s likelihood function to generate a set of posterior beliefs. Also most likely, the people of the third, second, and first millennia B.C. (and even substantially later) would not have cast the information-cum-belief-updating problem in these terms, but the structure itself is quite general and surely is age-old.¹⁰ Since the expert knows how the updating process works, she may be in a position to answer the question strategically or truthfully. The expert may have an interest in the decision that the client makes. Somewhat similarly, the expert may even have the best interests of her client in mind but believe that the client has either poor priors or an inaccurate likelihood function and may decide to offer an untruthful answer that the client will put into his likelihood function and so come up with the “best” or “right” decision despite his native shortcomings.

A client also would be aware of the potential for such strategizing. Several options are available for obtaining truthful information from the expert. First, the client could ask for the raw data, the expert’s likelihood function, or both. Asking for them and getting them may be two different things though. But, if the client can see the raw information, he can put it directly into his own likelihood function without the possibility of suffering distortions from having it passed through the expert’s likelihood function – and with a lower probability of getting deliberately false information, if only because it may be more costly to the expert to falsify the raw data than her own posterior beliefs – or because there may be some partial verifiability of the raw data. Transmission of the expert’s likelihood function may be more difficult: it could be tantamount to imparting an education, or it might involve the divulgence of arcane secrets in the cases of religious experts. Hiring multiple experts also could serve as a check on the candor of any individual

expert, particularly if each expert knows that others are being hired too.

The expert may or may not be induced to obtain the desired accuracy, even if such accuracy is within her capacity. Offering the expert a "cut" of any favorable consequences of the client's subsequent decision puts the expert in the position of maximizing her own wellbeing when she gives information that will maximize her client's wellbeing. This type of contract between client and expert can correct the kind of truthfulness problem that arises from outright conflicts of interest between expert and client, and it can *increase* accuracy but, as the expert still will not receive the entire benefit of increased accuracy, it may not maximize accuracy – unless the expert is offered the full consumer surplus.

Information and group decisions

When decisions must be made by multiple agents, as is common in many business and political settings, there are opportunities for disagreement that can affect their behavior in a number of ways. The decisions themselves can be affected, either through compromise or failure to reach a decision. Information acquisition strategies can be affected. The decision interactions themselves can be influenced by degrees and types of disagreement. Altogether, disagreements are more interesting than agreements – possibly the sort of reasoning that has reinforced economics' reputation as the dismal science. Disagreements in group decision making can emerge for two principal categories of reason, conflicts of interest and differences of opinion. We consider these two possibilities in turn.

Conflicts of interest can emerge from three major sources: differences in payoffs from any given decision, different utility functions for identical payoffs, and different endowments. A particular decision can yield different outcomes for different members of the decision council. For example, a decision to expand the navy would benefit the owner of a sizeable quantity of good timber land, but could threaten to raise the wage facing an urban employer who hires relatively unskilled labor (it depends on the size of the naval expansion relative to the size of the unskilled workforce). Some decision makers may

be more risk averse than others; some may have different preferences for items unrelated to the decision directly but affected indirectly by it. Finally, even if payoffs and utility functions were identical, if some of the decision makers had different distributions of income from sources other than those subject to the group decision, they may be affected differently. For example, the same action could be a gamble for one agent and an insurance provision for another. There is no simple way out of this problem by looking for a group preference scaling or a group probability estimate: they are generally impossible.

Differences in opinions are essentially differences in beliefs. These can be attributed to differences in priors or differences among likelihood functions (equivalent to different theories or models of the part of the world under discussion, certainly not an uncommon phenomenon today, even among people who would not want to be thought of as dealing with theories or models). Some of the individuals may also have private information, unavailable to others in the group.

The method of discussion also has economic implications. Suppose that the decision makers disclose their posterior beliefs rather than the raw information at their disposal. If the interacting agents are familiar enough with one another to have some understanding of one another's likelihood functions (that is, how each other views the world – or individual parts of it), they can work backwards from the revealed posterior beliefs expressed to at least an approximation to the initial information that was kept private.

Under some circumstances it may be rational for groups to agree not to seek certain information. In some uncertain situations, clarification of the uncertainty could introduce a conflict of interest that did not exist with the uncertainty. For example, suppose that a city of 5000 people decides to take a particular action to defend itself against an aggressor, on the understanding that 20 of them will probably be killed. This small, expected sacrifice may be acceptable to all the members of the city, but if it were possible to identify who the 20 sacrificees would be, the unanimity might dissolve. The same structure of potential conflict of interest characterizes many group decisions with probabilistic outcomes.

When everyone would benefit from an agreement that can be made only when outcomes are uncertain, it is irrational to seek the information that would prevent the agreement from being made.

In situations where differences of beliefs are multidimensional (that is, beliefs can differ on more than a single item of information), those differences can be offsetting in the sense that the combinations of multiple different beliefs lead to identical decisions. Suppose that we have two groups of officers looking at the city walls to decide where to reinforce them. They look at, say, zone A near the gate. The first group believes that it will rain, and because rain will especially impede visibility there, the Spartans are likely to press an attack there; consequently that zone needs reinforcing. The other group believes that it will not rain, but they also believe that the Spartans would be likely to attack that sector if the weather was dry; hence that area needs reinforcing. They reach the same decision on the basis of total disagreement on premises. If they invested more resources in resolving some of their differences, they might not be able to reach a decision on what to do to protect themselves from the Spartans.

In situations in which both conflicts of interest and differences of opinions exist, the decision makers will tend to avoid acquiring more information.

7.3 Dealing with Nature's Uncertainty

Early on in this chapter we noted that economic agents are confronted with two principal sources of uncertainty: that coming from the vagaries of nature and that coming from the strategic behavior of other agents, whether in markets or outside them. As complicated as the uncertainties coming from nature are, those coming from behavior are more intricate. This section eases us into those waters with the simpler problems of how economic agents deal with uncertainty emanating strictly from the state of nature. The first subsection picks up the threads of contingent consumption where we left them in section 7.1.3,

examining how an individual can arrange to “purchase” consumption in one state of the world or another before anyone knows which state will emerge. How can you contract to buy something that might not materialize yet still have the supplier willing to offer it? The answer involves assets, which may strike some readers as just about the farthest thing from what existed in antiquity as we have run across so far in this text. Naturally, we disagree, and we anticipate that by the time we’ve completed our exposition of the problem, readers will be anxious to apply the concepts to the ancient artifactual and textual data. The second subsection adds an additional layer of realism to the asset concept used in the examination of contingent consumption by allowing interactions between the assets, interactions that we will call covariances. These interactions are a type of risk, but in some cases they can actually reduce risk in a collection of assets called a portfolio.

7.3.1 *Contingent markets*

We have already introduced the concepts of contingent consumption – consumption of different bundles of goods under different states of the world that might emerge in the future – and state claims, which are contracts that give the holder the right to consume these contingent consumption bundles. The different states of the world could be good weather versus bad weather, war versus peace, earthquake versus no earthquake, good flood versus bad flood, head of household gets sick and can’t work versus head of household stays healthy. Basically, you name it. We have given the name “contract” to the understanding (oral or written) that entitles a person to consume specified bundles of commodities in different states of the world, but what one of these contracts looks like may remain vague. It is related to a “price tag,” in the sense that it is a guarantee of a particular amount of consumption under specified conditions in the future, and the future consumption ratio implicitly defines a relative price or even a set of relative prices. Yet this “future price tag” agreement itself has a price tag on it because it is obtained only in exchange for an amount of resources specified today.

State-claims and assets

We will study the provision of these contracts through two mechanisms, exchange in a market and self-provision. These contracts take the form of assets that are purchased now for use in the future. The state-claims we studied in section 7.1.3 were entitlements to consume a stated amount of “stuff” (a specified commodity bundle in more clinical terminology) in the event state s occurred in some future period. But what if state $s + 1$ occurs? Do we want to just be out of luck in that case? Undoubtedly not. It would be useful to have the opportunity to guarantee our consumption at one level if state s occurs and some other level, possibly higher or lower, if state $s + 1$ should occur. It would be even more useful if we could arrange for this set of contingencies with a single agreement. A contract can be written to accomplish just that, and that contract takes the form of an asset that entitles the holder to either specified or anticipated consumption levels in alternative states of the world. In the contemporary world, stocks in publicly traded companies serve this role. A share of a particular stock will pay us an amount x in one state of the world, an amount y in another state of the world, and an amount z in yet another state of the world. Typically these amounts to be paid in particular states of the world are not specified, but traders’ experience with a large number of such stocks over time has informed them with some degree of accuracy what kinds of returns can be expected in particular conditions. Think about the reputations of, say, utility stocks (gas, water and electric power companies), “industrials” (manufacturing companies), and services (retail companies) over the business cycle: utilities typically have more stable returns, industrials fluctuate quite a bit, and services are somewhere in between.

Let’s take this asset concept to a more home-spun level, say the ownership of a farm and the ownership of a small, family business. An individual may jointly own either of these physical assets or may own well specified shares entitling him to a specific percentage of the net income from them. The farm can be expected to prosper according to, among other events, the weather in particular. The business in town may not be so strongly affected by the weather, although if most of the clientele derive their income from

farms, the town and farm incomes certainly will tend to move together. If the town business sells its products to a wider area – say it is a genuine “export” business – the contingent consumption its income offers may vary less among local states of the world than would that conferred by farm income. Foreign war or foreign peace might affect the business considerably and the farm not at all. Both the farm and the business venture are physical assets; shares of ownership in either would be financial assets, whether those shares are codified or simply understood. An ownership share in either offers its holder the prospects of particular consumption levels in different states of the world.

Maximizing state-contingent utility

The price that a share in either the farm or the business would command is related to the underlying state-contingent incomes each provides. We develop this particular relationship now. Consider a situation in which two states of the world, $s = 1, 2$, are possible and in which there are two assets, $a = 1, 2$, with prices p_1 and p_2 . We use a wealth constraint to show the limits on asset holdings: $p_1 q_1 + p_2 q_2 = p_1 \bar{q}_1 + p_2 \bar{q}_2 = \bar{W}$, where the q_i are the quantities of each asset, W is wealth, and a bar over a q_i and W represents initial endowments. Thus, we start with a situation in which the individual with this budget constraint possesses the amounts $\bar{q}_1$ and $\bar{q}_2$ of assets 1 and 2. Define the income, or the per period return, from each asset a in each state s as z_{as} . With this notation, z_{12} would be the income from asset 1 if state 2 occurs, z_{21} would be the income from asset 2 if state 1 occurs, and so on. Then, with the two assets and the two states, the consumption our individual could get in state 1 would be $c_1 = z_{11}q_1 + z_{21}q_2$ (we’re assuming that the unit prices of contingent consumption bundles c_1 and c_2 are both 1), and that in state 2 would be $c_2 = z_{12}q_1 + z_{22}q_2$. Let’s develop a bit more notation to simplify the next few steps. Let the share of wealth allocated to each asset be $k_i = p_i q_i / \bar{W}$. Then we can rewrite these two expressions for contingent consumption in terms of the fractions of wealth allocated to the assets conferring consumption rights: $c_1 = k_1 z_{11}(\bar{W}/p_1) + k_2 z_{21}(\bar{W}/p_1)$ and $c_2 = k_1 z_{12}(\bar{W}/p_1) + k_2 z_{22}(\bar{W}/p_2)$. Now we can

set up the individual's expected-utility maximization problem as $\text{Max}_{k_1, k_2} L = \pi_1 v(c_1) + \pi_2 v(c_2) - \lambda(k_1 - k_2 - 1)$, in which we have substituted k_1 and k_2 in the budget constraint. (Recognizing that $k_2 = 1 - k_1$, the budget constraint is equivalent to $2k_1 - 2$, and by adjusting k_1 so as to maximize utility deriving from asset 1 we are implicitly adjusting k_2 as well.) The change in contingent consumption in state s from adjusting either asset share k_a has the relationship $\Delta c_s / \Delta k_a = z_{as} (\bar{W} / p_a)$. Then, rearranging the first-order conditions from the maximization we get the relationship between asset prices, asset quantities, and expected marginal utilities of contingent consumption in the two states represented by $(\sum_{s=1}^2 \pi_s v'(c_s) z_{1s}) / p_1 = (\sum_{s=1}^2 \pi_s v'(c_s) z_{2s}) / p_2$, in which the notation v' is a compact expression for $\Delta v / \Delta c_s$. This relationship gives the individual the same marginal expected utility per unit value (dollar, franc, shekel, drachma) held in each asset. This relationship can be generalized to a modification of the Fundamental Theorem of Risk-Bearing developed in section 7.1.3: $(\sum_s \pi_s v'(c_s) z_{1s}) / p_1 = (\sum_s \pi_s v'(c_s) z_{2s}) / p_2 = \dots = (\sum_s \pi_s v'(c_s) z_{as}) / p_a$, for all assets $a = 1$ through a and all states s . The numerators in the expression of the previous sentence are the sums across states of the expected marginal values of an additional unit of the asset; the denominators are their marginal costs. The $\pi_s v'(c_s)$ terms in the numerators are the expected additional utility of increasing contingent consumption c_s by a small amount in state s ; the expectation is described by the particular probability a consumer assigns to the π_s term. The z_{as} term is the additional amount of purchasing power that can be spent on the contingent consumption good c_s with an additional unit of asset a .

Now let's consider the relationship between the prices of the assets (p_a) and the prices of the underlying state-claims (p_s) and the question whether trading in assets is equivalent to direct trading in state-claims (the c_s). Asset prices are related to contingent-claim prices as the sum of the products of the income from the asset in each state and the contingent claim price for that state: $p_a = \sum_s z_{as} p_s$. If the array of assets that can be traded (exchanged) constitutes a complete set of contingent markets, trades in assets will replicate the results of trades in state-claims. If the

assets represent an incomplete set of contingent markets, the results would not be the same. What makes for a complete or an incomplete set of contingent markets? In general, if the number of assets is greater than or equal to the number of possible states, we may have complete contingent markets.¹¹ If $A \geq S$ and at least S of the A assets have linearly independent yield vectors (that is, some of the assets' yields aren't just multiples or slightly more complicated relationships of other assets' yields), we have a complete set of asset markets. We can always calculate the asset prices if we know the state-claim prices; if there are at least S linearly independent assets, we also can calculate the state-claim prices from the asset prices.

Complete markets

In a market equilibrium (actually, a complete asset market equilibrium), with all individuals facing given asset prices, all individuals' ratios of expected marginal utilities would be equalized and would themselves equal the ratio of state-claim prices: $[\pi_{sn}^j v'_j(c_{sn}^j)] / [\pi_{sm}^j v'_j(c_{sm}^j)] = [\pi_{sn}^k v'_k(c_{sn}^k) / \pi_{sm}^k v'_k(c_{sm}^k)] = p_{sn} / p_{sm}$ for all individuals j and k , for all pairs of states $s = n$ and $s = m$. This notation may appear especially grueling, but let's take it a bit at a time. Each term is a ratio of expected marginal valuations from putting a bit more of one's wealth into alternative assets. At the margin, one would want to get equal additions to one's wellbeing from any given allocation. The π_s^j terms are just individual j 's personal assessment of the probability of state s occurring; the v'_j are individual j 's marginal utility (technically, marginal preference scaling) of a small increment of contingent consumption in state s , c_s^j . Each ratio compares the expected marginal valuation of the additional contingent consumption expected to derive from an additional drachma (or pair of chickens, bag of lentils, or whatever unit of barter individuals use) in asset a , relative to the comparable marginal valuation from an extra drachma in some other asset. The second such ratio presents the same set of conditions for individual k ; we could have gone on to show the same sets of ratios for many more individuals, but they all would have looked the same. Note that if individuals have different probability beliefs π_s , the ratios of marginal utilities (or to be quite

specific, marginal state preference orderings) need not be the same.

In addition to such equalized marginal conditions, the sum of the desired holdings of state claims across all the individuals in the market must equal the sum of the endowments with which they started: $\sum_j c_s^j = \sum_j \bar{c}_s^j$. And contingent consumption in each state must be equal to the sum over assets of the state-contingent incomes yielded by each asset: $c_s = \sum_a q_a z_{as}$, for each state s . This insures that the market is not trying to allocate either more or less than the amount of contingent consumption available in each state. The individual ratios of marginal utilities from incremental asset holdings also have a market-wide equivalent to that for the individual state-consumption marginal utilities: $[\sum_s \pi_s^j v'_j(c_s^j) z_{as}] / [\sum_s \pi_s^j v'_j(c_s^j) z_{is}] = [\sum_s \pi_s^k v'_k(c_s^k) z_{as}] / [\sum_s \pi_s^k v'_k(c_s^k) z_{is}] = p_a / p_i$. The clearing condition for the asset market is that the number of shares desired (demanded) equal the number endowed: $\sum_j q_a^j = \sum_j \bar{q}_a^j$, for all assets $a = 1$ through n .

Incomplete markets

If there are not at least S linearly independent assets, we are in an incomplete-markets world, and trading in asset markets is not equivalent to direct trading in state claims. The Risk-Bearing Theorem for Assets will always “hold” (technical lexicon – jargon – for the fact that the equalizations of ratios in that theorem can still occur), but the Fundamental Theorem of Risk-Bearing will not, because there are not separate vehicles (assets) with which to transact for each contingent claim (state claim). Consequently there is not enough “flexibility” in the linked chain of ratios for all of them to be equalized. In such a case, transactions for several contingent claims might be “packaged together” and sold as a fixed bundle, whether individuals want to purchase the particular ratio of the items inside them or not. For instance, short-term consumption loans might be packaged together with land rentals, and renters will be paying an amount for their land that includes some charge for the consumption loans that need not equal the going loan rate times the loan they take out (which might, in some cases, be zero). Such a situation is characterized in the descriptive literature as tied loans or, somewhat more poignantly, debt bondage or debt peonage.

Such incompleteness could occur if, simply, there are fewer assets than states, if there are as many assets as states but not all the assets are linearly independent, or if some of the assets are not tradable, all of which are conditions which are interesting in the contexts of the economies of the ancient Mediterranean and Aegean. A particularly important asset for many individuals is their own future labor earnings. Today, slavery and indenture are generally outlawed, a fact that, despite all the freedoms those prohibitions confer, restricts individuals’ abilities to diversify their consumption. Americans in particular may be accustomed to thinking of slavery as a largely irreversible relationship: in general, once a slave, always a slave. Slavery in the ancient Mediterranean world, while frequently irreversible and equally frequently harsh, did, at times and in some places, provide for temporary slavery which was equivalent to bonding one’s labor for a specified period in return for current consumption.

Production adjustments for contingent consumption

So far we’ve considered market exchanges as the only means of providing oneself with the array of contingent consumption one wants. Now we introduce the use of production to convert one’s endowment of contingent claims to a more desired, but still attainable, set. We use an example from agricultural production. Suppose an individual farmer faces two states of nature: in state 1, adequate rainfall occurs and the crop is good; in state 2 the weather is dry and the crop fails completely. In Figure 7.17, we measure consumption c_s in states 1 and 2. A transformation curve shows how the farmer can convert consumption in the good-weather period into consumption in the bad-weather period. If he were simply to get the most out of the good season and take no care for the fact that he will starve if the other state occurs, he will produce at point $\bar{c}_1$ on the horizontal axis, which has the coordinates in c_1 – c_2 space of $(\bar{c}_1, 0)$. However, he can direct some of his effort away from his routine cultivation during good weather to activities such as terracing and developing irrigation facilities that will raise his crop yield above zero should state 2 occur. In doing this, he sacrifices some of the potential state-1 output he could

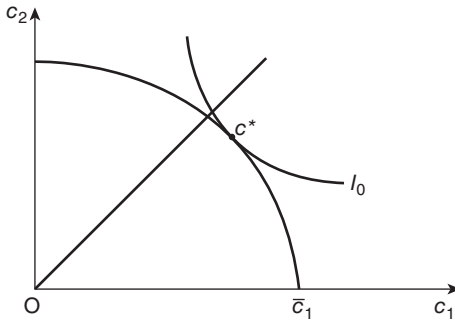


Figure 7.17 Transforming the outcomes of a risky event.

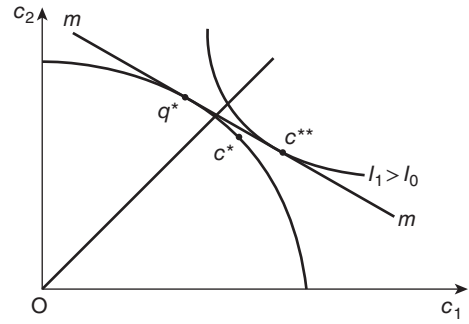


Figure 7.18 Using the market to transform the outcomes of a risky event.

achieve. Now draw the 45° certainty line. His preferences for state-contingent consumption are represented by the indifference curve tangent to the transformation curve just to the right of the certainty line at c^* .

The general formulation for the transformation curve (the opportunity constraint) is $f(q_1, q_2) = 0$. Then the tangency of the indifference and transformation curves is characterized by the conditions $f_1/f_2 = -\Delta q_2/\Delta q_1|_f = -\Delta c_2/\Delta c_1|_u \equiv \pi_1 v'(c_1)/\pi_2 v'(c_2)$. The notation $|_f$ indicates that the change in q_2 per unit of change in q_1 (the marginal rate of transformation between output in state 1 and output in state 2) takes place along the transformation curve defined by the function f and $|_u$ indicates that the change in state-2 consumption per unit of state-1 consumption (the marginal rate of substitution between consumption in the two states) takes place along the indifference curve defined by the utility function u ; and $f_i \equiv \Delta f/\Delta q_i$. The final term, with the ratio of probabilities of states 1 and 2, indicates that the tangency obeys the fundamental theorem of risk-bearing. In general, the producer will not want to avoid all risk, which would be equivalent to a tangency at the intersection of the certainty line with the transformation curve, even though it may be possible to do so.

Now suppose that the farmer can use both productive adaptations and the market to achieve the contingent consumption that maximizes his utility. In Figure 7.18, we have the same transformation curve, with a market-determined relative price of state-2 to state-1 consumption denoted by mm , tangent to the transformation

curve at q^* . Indifference curve I_1 is from the same family of indifference curves as I_0 in Figure 7.17, but is farther from the origin and hence represents a higher level of utility. Point c^* on the transformation curve indicates the original tangency when there were no market opportunities for exchange of state claims. Now, the indifference curve I_1 is tangent to the market price line at c^{**} , offering the farmer slightly less consumption in state 2 but considerably more in state 1. However, he will organize his productive activities to produce at q^* on his transformation curve – the tangency of the market price line with his transformation curve – which has him producing much less in state 1 but considerably more in state 2. The existence of the market opportunity has let him separate his production risk-bearing from his risk-bearing in consumption. We have to reformulate the equalization of marginal conditions as $-\Delta q_2/\Delta q_1|_f = p_{s1}/p_{s2} = -\Delta c_2/\Delta c_1|_u$ to include the role of the state-claim prices.

State-dependent utility

So far we have considered the utility function to be invariant among states of the world – it stays the same regardless what state of the world occurs. Utility is a function only of consumption levels. Of course, utility may be a function of what state of the world occurs as well as the consumption level that is possible. Suppose one state of the world was that all your family was still alive and the other state was that all of them but yourself were dead. The utility one would derive from identical consumption levels in the two states very likely would differ. Since this

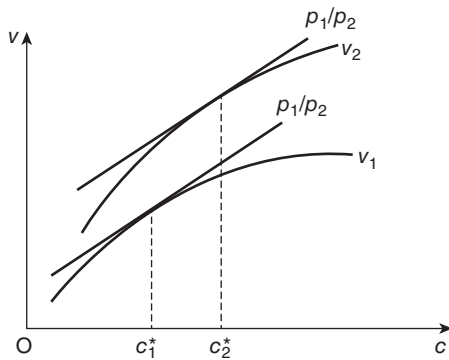


Figure 7.19 Equilibrium in risky consumption choices.

specification still appeals to a single, underlying preference function, the same results from using the expected utility rule will apply. The fundamental theorem of risk-bearing yields the relationship $\pi_1 v'_1(c_1)/p_1 = \pi_2 v'_2(c_2)/p_2$, or $[\pi_1 v'_1(c_1)]/[\pi_2 v'_2(c_2)] = p_1/p_2$. In this case, the notation v'_s indicates the state- s marginal utility (preference scaling) of an additional bit of consumption in state s . Once again these expected marginal utility ratios are equated to the ratio of state-claim prices (state-contingent consumption prices). Figure 7.19 shows the preference scaling function as a function of state-contingent consumption c , for each state: v_1 for state 1 and v_2 for state 2. The two price lines p_1/p_2 are parallel, representing the same relative state-claim price, but they yield widely divergent, optimal consumption. The concept can be used to study insurance behavior, but I do not pursue the analysis further here.

7.3.2 Portfolios and diversification

A group of assets, whether financial claims as is common in today's industrial societies – at least in the upper reaches of the personal income distribution – or physical assets such as land, buildings, and equipment, forms a portfolio. Portfolios have interesting properties that are composed of, but do not always mirror, the properties of the individual assets composing

them. Someone holding a portfolio will be interested in two principal characteristics of it: its expected rate of return and the riskiness of that expected return. In this section, we will show how these properties of a portfolio are built up from the corresponding properties of individual assets; the characteristics of an efficient portfolio from the perspective of the individual; and how the aggregate operation of a capital market prices these assets.¹² We will show two especially far-reaching facts about risk in this subsection. First, the risk of an asset, be that asset a share of common stock in General Motors or a plot of land planted in chickpeas, depends on whatever else the owner of the asset owns. What looks on the face of it to be highly risky may actually reduce risk. Second, the effect of an asset's riskiness on its price depends on the riskiness of everything else, but not at all on the current use of the asset.

The individual portfolio

Let's start with the definition of the rate of return, so we know exactly what we are describing. We denote the rate of return on an asset i in some period t as R_{it} . The total rate of return is composed of any current-period income (a dividend) plus the capital gain from any price change in the asset over the period. We have $R_{it} = (d_{it}/p_{i,t-1}) + [(p_{it} - p_{i,t-1})/p_{i,t-1}]$; d_{it} is the income (dividend) from one unit of the asset between the end of period $t-1$ and the end of period t (or over period t); p_{it} is the price (market value), at the end of period t , of a unit of the asset, purchased at the end of period $t-1$; and $p_{i,t-1}$ is the price of a unit of the asset at the end of period $t-1$. Keep in mind that R_{it} measures rates of return, which are usually relatively small fractions, roughly the magnitude of interest rates, not absolute incomes per period.

We denote the expected return on an asset i $E(\tilde{R}_i)$, where the tilde ($\sim$) indicates a random variable – that is, one that takes on uncertain values and consequently is risky. We use the term x_i to represent the share of asset i in a portfolio: $x_i = p_i s_i / \sum_i p_i s_i$, where s_i is the quantity of asset i and $\sum_i x_i = 1$. Using these symbols, the expected rate of return on a portfolio of

assets is, quite simply, $\sum_{i=1}^n x_i E(\tilde{R}_i)$. For the variance of the portfolio's expected return, think back to section 7.2.1 to the definition of a variance (denoted by σ^2), and think about the structure of the expected return of a portfolio. The variance is $\sigma^2(\tilde{R}_p) = E\{[\tilde{R}_p - E(\tilde{R}_p)]^2\}$. Substituting from the definition of the portfolio's expected return, this is equivalent to $\sigma^2(\tilde{R}_p) = E([\sum_i x_i (\tilde{R}_i - E(\tilde{R}_i))]^2)$, which you will notice is quickly taking on a quadratic form: recall from eighth-grade algebra expressions like $(x + y) \cdot (x + y)$, or $(x + y)^2$, which, when multiplied out, equaled $x^2 + 2xy + y^2$. We will next rewrite the expression for the portfolio variance in a form that will look like this: $\sigma^2(R_p) = \sum_i x_i \sigma^2(R_i) + \sum_i \sum_{j \neq i} x_i x_j \sigma_{ij}$, or $= \sum_i \sum_j x_i x_j \sigma_{ij}$, where σ_{ij} is the covariance between the returns to assets i and j (σ_{ii} would be what we have referred to with the symbol σ^2 , the variance; the definition of the covariance between two random variables is $\text{cov}(X, Y) = 1/n \cdot \sum_i (\bar{X} - X_i)(\bar{Y} - Y_i)$). In this last expression, the double summation $\sum_i \sum_j x_i x_j = 1$. Where x_i represented the share of the portfolio's value held in asset i , x_j is the similar representation of the share of the portfolio held in asset j , where we use the subscript designations i and j simply to represent different assets. Finally, we can express the variance of a portfolio as the weighted sum of the "covariance risks" of the individual assets: $\sigma^2(\tilde{R}_p) = \sum_i x_i \text{cov}(\tilde{R}_i, \tilde{R}_p)$. In the next to last expression for the portfolio variance, we divided the variance into two components, the sum of direct variances of the individual assets and the sum of the covariances among the assets. Thus the variance of a portfolio's rate of return could be greater or less than the sum of the individual assets' variances.

The risk of any security in this portfolio is $x_i \sigma_{ii} + \sum_j x_j \sigma_{ij}$, where j runs from asset 1 to asset n but excludes asset i , or simply $\sum_j x_j \sigma_{ij} = \text{cov}(\tilde{R}_i, \tilde{R}_p)$. The risk of any security relative to the risk of the entire portfolio can be characterized by what is termed "the beta" of the security: $\beta = \text{cov}(\tilde{R}_i, \tilde{R}_p) / \sigma^2(\tilde{R}_p)$, the covariance between the individual asset and the entire portfolio, divided by the variance of the portfolio return. If $\beta_i > 1$, the risk of security i is greater

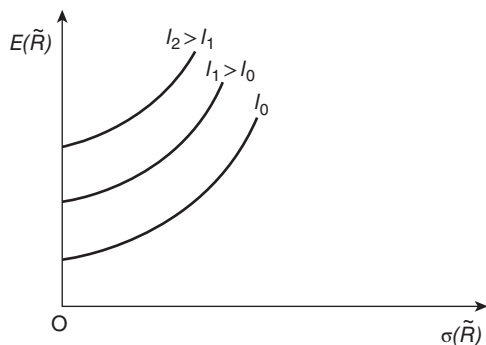


Figure 7.20 Indifference curves for risk and expected rate of return.

than the overall portfolio risk; it adds risk to the portfolio; if $\beta_i < 1$, it is less risky.

Recall that the portfolio holders are interested in the expected return and the risk of that return; and recall the mean-variance risk model presented in Chapter 3; I take the opportunity of Figure 7.20 to refresh your recollection of that model. The vertical axis is the mean, or the expected value, of the variable in question, the rate of return in the particular case at hand. The standard deviation is measured on the horizontal axis; remember that the standard deviation, represented by σ , is just the square root of the variance σ^2 . The indifference curves I_0 , I_1 , and I_2 show the combinations of the expected rate of return and its risk, as measured by the standard deviation of that rate of return, that will leave an individual indifferent. If the risk increases, the consumer must be compensated with a higher expected rate of return to make him as well off. Similarly, if the expected rate of return is higher, the individual will be willing to accept more risk and still feel equally well off. Consequently these indifference curves indicate increasing well-being as we move from southeast to northwest; their convexity represents risk aversion. We will develop what is called "the efficient set" of portfolios in the same "mean-variance space" as this picture of consumption preferences, measuring the mean by the expected value of the entire portfolio and the risk of the portfolio by the standard deviation of its overall rate of return.

We develop the algebraic and graphical expressions for the efficient set of portfolios using the simplest possible portfolio: a combination of two risky assets we call asset i and asset j . The rate of return on this portfolio is $\tilde{R}_p = x\tilde{R}_i + (1-x)\tilde{R}_j$, its expected rate of return (looking to the future, which is currently unknown, rather than looking at already realized values) is $E(\tilde{R}_p) = xE(\tilde{R}_i) + (1-x)E(\tilde{R}_j)$, just the share-weighted sum of the expected rates of return of the two assets. Using the expression for the variance of the portfolio return developed in the previous paragraphs, but for only these two assets, the standard deviation of the portfolio return is $\sigma(\tilde{R}_p) = [x^2\sigma^2(\tilde{R}_i) + (1-x)^2\sigma^2(\tilde{R}_j) + 2x(1-x)\text{cov}(\tilde{R}_i, \tilde{R}_j)]^{1/2}$, which has exactly the quadratic form of $[(x+y)^2]^{1/2}$. It will help in the connection between the symbolic and pictorial representations to convert the covariances into correlations, which can take values between plus and minus one: $\text{corr}(\tilde{R}_i, \tilde{R}_j) = \text{cov}(\tilde{R}_i, \tilde{R}_j)/\sigma(\tilde{R}_i)\sigma(\tilde{R}_j)$, or simply the covariance divided by the product of the respective standard deviations. Substituting this expression for the correlation between the returns on assets i and j , we have $\sigma(\tilde{R}_p) = [x^2\sigma^2(\tilde{R}_i) + (1-x)\sigma^2(\tilde{R}_j) + 2x\text{corr}(\tilde{R}_i, \tilde{R}_j)\sigma(\tilde{R}_i)\sigma(\tilde{R}_j)]^{1/2}$. The advantage of this latter expression will become apparent as we develop the geometric representation of the standard deviation of portfolio return in Figure 7.21.

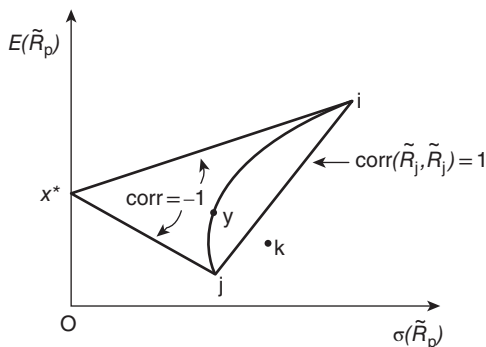


Figure 7.21 An efficient portfolio of risky assets with different risks and expected returns.

In Figure 7.21, points i and j represent the coordinates of expected return and standard deviation for assets i and j . If the correlation between the returns on the two assets is plus one, the standard deviation of the portfolio return is just the weighted sum of the individual asset standard deviations: $\sigma(\tilde{R}_p) = x\sigma(\tilde{R}_i) + (1-x)\sigma(\tilde{R}_j)$. For given portfolio shares, x and $1-x$, and asset return variances, the variance of the portfolio is highest when there is a perfect correlation (plus 1) between the asset rates of return. Diversification, accomplished by varying the asset share x , is totally ineffective in reducing portfolio risk. In Figure 7.21, all the combinations of asset i and asset j in the portfolio composed of the two of them are represented by the straight line between them. As we move between points i and j on that line, we are changing the portfolio shares x and $1-x$; when $x = 1$ and $1-x = 0$, we are at point i ; when $x = 0$ and $1-x = 1$, we are at point j . No combination of assets in the portfolio above line ij is possible, and no combination of assets below it is efficient. With any combination of assets represented by a point below the line, for instance point k , we could get either lower risk at the same expected rate of return or a higher expected rate of return for no more risk – or a combination of lower risk and higher expected rate of return.

The correlation between the returns on assets i and j has no effect on the expected value of the portfolio, but it has considerable influence on the standard deviation of the portfolio return, which we will see when we look at the shape of the efficient portfolio when the correlation is minus one. In this case the portfolio risk is the difference between the individual, weighted asset return risks: $\sigma(\tilde{R}_p) = x\sigma(\tilde{R}_i) - (1-x)\sigma(\tilde{R}_j)$. It is easy to see that the right-hand side of this expression could go negative, but we know that a standard deviation is always positive. When the portfolio share of asset i is $x_i = \sigma(\tilde{R}_i)/[\sigma(\tilde{R}_i) + \sigma(\tilde{R}_j)]$, the standard deviation of the portfolio return is exactly zero, represented by point x^* on the vertical axis of Figure 7.21. From $x = 1$ to $x = x^*$, the efficient set is represented by the straight line from point j to x^* . At that point, and for values of x smaller than x^* , we have to turn the terms in the right-hand side of the expression around and

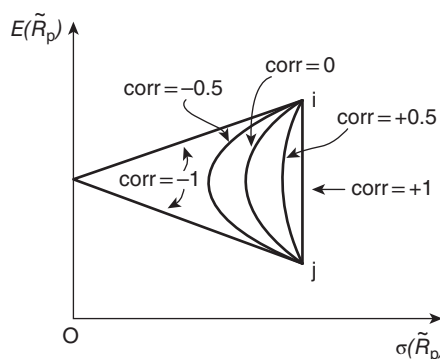


Figure 7.22 Efficient portfolios under alternative correlations between assets' expected returns.

subtract the share-weighted standard deviation of asset i from the share-weighted standard deviation of asset j . This is represented by the straight line from x^* to point j .

For correlations intermediate between 1 and -1 , the efficient set is curved toward the vertical axis (its shape is a parabola: recall the Cartesian graph of a quadratic formula as you drew them in the eighth grade). One example is shown by the line jy_i in Figure 7.21. Any point on that line to the southeast of y on that efficient set is in fact inefficient: the same risk could be achieved with a higher expected return on the upper half of the parabola. Figure 7.22 shows the different degrees of bowing associated with different correlations between the assets' expected rates of return.

Sometimes it is possible to combine a riskless asset with a portfolio of risky assets. In such a case, the rate of return on the total portfolio is $\tilde{R}_p = xR_F + (1-x)\tilde{R}_r$, where R_F is the rate of return on the risk-free asset (which is why R_F does not have a tilde over it), x is the share of the portfolio in that asset, and $\tilde{R}_r$ is the rate of return on the risky portion of the portfolio. The standard deviation of the total portfolio then is just the standard deviation of the risky portion of the portfolio times its share in the portfolio: $\sigma(\tilde{R}_p) = (1-x)\sigma(\tilde{R}_r)$. In the efficient portfolio diagram, the efficient portfolio becomes the straight line between the risk-free rate of return on the vertical axis and a tangent to that line with

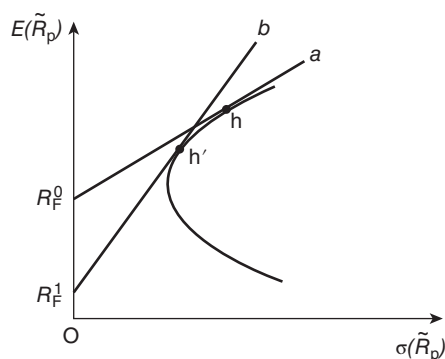


Figure 7.23 Portfolio composition when a risk-free asset is available.

the efficient risky portfolio parabola, as drawn in Figure 7.23. With a risk-free rate of R_F^0 , the efficient set is the segment of the straight line $R_F^0 - a$ between R_F^0 and h .¹³ The total portfolio would be composed of fraction x of the risk-free asset and $1-x$ of the risky portfolio defined by point h . With a lower risk-free rate, such as R_F^1 , the efficient set would extend along the line $R_F^1 - b$ from R_F^1 to point h' on the risky portfolio parabola, indicating that the risky portion of the efficient portfolio would consist of lower-return, less risky assets.

This exposition has shown that an apparently "risky" asset ($\sigma > 0$) may have a positive, zero, or even negative risk in a particular portfolio. We can show an individual asset's contribution to the risk of a portfolio several alternative ways, each showing a different aspect of the relationship. First, we have $x_i(\sum_j x_j \sigma_{ij})$, where the summed term in parentheses is the weighted, pairwise covariance between asset i and each other asset j . This risk is weighted by the share of the asset in the portfolio. Alternatively, if we consider asset i one of the j assets (one of the covariances σ_{ij} is the variance σ_{ii}), we can rewrite the sum of covariances expression as $\sum_j x_j \sigma_{ij} = x_i \sigma_{ii} + \sum_{j \neq i} x_j \sigma_{ij}$, which emphasizes that one part of the risk that an asset contributes to a portfolio is proportional to the variance of its own rate of return and another, possibly even larger, component is its covariance with all other

assets in the portfolio; the proportionality factor is the distaste for risk relative to the preference for expected return. It is important to note that the σ_{ij} relationships are invariant to the assets in a particular portfolio – these covariances exist between assets whether they are held in the same portfolios or not. How the covariances affect the risk of a particular portfolio depends on which assets are combined in the portfolio. Any particular asset can make different contributions to risk in different portfolios. This last expression for the risk that an asset contributes to a portfolio also suggests that, if a portfolio were large enough, the variance of an asset's own rate of return might make virtually no contribution to portfolio risk. Experience indicates that by the time 15 to 20 assets are in a portfolio, the only source of risk from any individual asset is its covariance risk.

Let's think about the portfolios of, say, farmers in any of the ancient Mediterranean or Aegean regions. How could we think of their "assets"? First, their principal source of wealth was their own labor. Allocating their labor to different activities would be equivalent to putting various shares of their wealth into particular assets. Different crops, animal husbandry, part-time craft activities, a part-time "job" in town rather than on the farm, even *corvée* labor provided it was remunerated at least in terms of subsistence meals. The part-time town job or the *corvée* labor might function roughly as "risk-free" assets if their "wages" were fixed and known in advance. Take another look at Figure 7.23. Let's think of R_F^0 and R_F^1 as the risk-free rates of return implied by the same town wage rate paid to farmers working part-time off-farm but living at different locations. The farmer living farther from the town employment receives the equivalent of R_F^1 since he has to debit his transportation costs (assume they both travel as frequently between home and town, but one travels farther). The one living farther from town would combine risk-free town labor with a less risky combination of crops on the farm. Riskier crops would be grown closer to town if farmers have part-time work in town available to them. Note that they need not *take* the town work for their risky farm portfolios to be affected by the available wage work: $x = 0$ is a possibility, in which case all their time (their wealth) would be applied to their risky assets.

Market equilibrium and the "capitalization" of risk in asset prices

We will develop two principal relationships in this subsection. First, we will show how a risk premium in an asset's rate of return emerges from the optimization of assets in a portfolio. Following that, we will turn to the full market and show how the risk of an asset is incorporated into its price as a result of a large number of individual portfolio holders' optimization decisions.

The portfolio optimization can be accomplished by minimizing the variance of a given portfolio of assets – through adjusting the asset shares x_i – subject to getting a minimum expected rate of return out of the portfolio and allocating the full value of available wealth among the assets. This problem is: $\text{Min}_{x_i} \sigma^2(\tilde{R}_p)$, subject to (i) $\sum_i x_i E(\tilde{R}_i) = E(\tilde{R}_e)$, in which $E(\tilde{R}_e)$ is some given level of expected portfolio return, and (ii) $\sum_i x_i = 1$. We form the Lagrangean (which I won't show; remember the Lagrangean for an optimization problem – it's the objective function, the expression to be maximized or minimized, plus a multiplier times the first constraint, plus another multiplier times the second constraint) and take the first-order conditions by adjusting the x_i and the two multipliers so as to minimize the value of the Lagrangean. From this procedure, after some rearrangements, we get an expression that relates the expected return on any (each) asset i and the risk of that asset in the portfolio that gets expected return $\tilde{R}_e$. Specifically, the difference between the expected return on any asset i and the expected return on the portfolio is proportional to the difference in the risk of asset i in the portfolio and the risk of the portfolio: $E(\tilde{R}_i) - E(\tilde{R}_e) = [\sum_j x_j \sigma_{ij} - \sigma^2(\tilde{R}_e)]/\lambda$, in which λ is the Lagrange multiplier on the expected return constraint. Some further rearrangements yield a decomposition of the expected rate of return on each asset i : $E(\tilde{R}_i) = E(\tilde{R}_{0e}) + [E(\tilde{R}_e) - E(\tilde{R}_{0e})]\beta_i$. The term β_i has the same interpretation it had just above: the covariance between asset i 's return and the return on the portfolio, divided by the variance of the portfolio return. The expected return $E(\tilde{R}_{0e})$ is the expected return on any asset whose return is uncorrelated (correlation equals zero) with the return on the entire portfolio. This expression says that the return on any asset is equal to the return on an asset that is riskless in the portfolio plus a risk premium equal to β_i times

the difference between the expected return on the total portfolio and that on the risk-free asset; the beta coefficient makes this risk premium specific to asset i , since it contains the covariance between asset i and the portfolio. As might be expected, something quite close to this risk premium shows up in the capital asset pricing formulation, which we take up next.

Now, to develop the market equilibrium, we consider a representative investor / asset holder who maximizes his welfare as a function of his certain consumption in the present period, his expected future consumption and the variance of his expected future consumption, subject to a wealth constraint which says that the sum of his current-period investments and his current-period consumption can't exceed his current wealth: $\text{Max } F_1(c_{1i}, E_1\sigma_i^2)$, subject to $w_{1i} = \sum_j x_{ij}p_j + c_{1i}$. Investor i 's current consumption is c_{1i} , his expected future consumption is $e_i = E(c_{2i})$, where the subscript 2 indicates "period 2"; the variance of current consumption is $\sigma_i^2 = \sigma^2(c_{2i})$, w_{1i} is current period wealth, x_{ij} is the share of firm j 's assets that individual i owns, and p_j is the current value of the firm.

As usual, we form the Lagrangean – which we again do not show – and take the first-order conditions by performing the usual adjustments of current consumption c_{1i} , current asset ownership shares x_{ij} and the Lagrange multiplier. With the customary rearrangements of these first-order expressions, we obtain an expression for the individual investor's demand function for the shares of assets of each firm: $2\sum_k x_{ik}\text{cov}(\tilde{V}_j, \tilde{V}_k) = p_j(\Delta\sigma_i^2/\Delta c_{1i}) + E(\tilde{V}_j)(\Delta\sigma_i^2/\Delta E_i)$. The $E(\tilde{V}_j)$ term is the expected value of the firm at time 2, and the covariance term on the left-hand side of the demand expression is the covariance between the expected values of different firms. The $\Delta\sigma_i^2/\Delta E_i$ term is the marginal rate of substitution between the mean and variance of future consumption, and $\Delta\sigma_i^2/\Delta c_{1i}$ is the corresponding marginal rate of substitution between current consumption and the variance of future consumption; both MRS terms are negative. To find the value of any firm j in a market equilibrium of all asset holders, we add the demand functions across all the asset holders and require that $\sum_i x_{ij} = 1$ for each firm j (which just says that the full value of each firm is held among the full set of asset holders); with a few more

rearrangements, we get an expression for the value of the firm (or value of an asset) that has the form $p_j = [E(\tilde{V}_j) - \theta\text{cov}(\tilde{V}_j, \tilde{V}_m)]/[1 + E(\tilde{R}_{0m})]$, in which $E(\tilde{V}_m)$ is the expected value of the market portfolio, and θ is an intricate expression describing the difference between the expected aggregate future value of all firms (assets) and their current value, divided by the variance of the expected aggregate value; this term is, essentially the price of risk reduction. The capital asset pricing rule, known as the Sharpe–Lintner capital asset pricing model (CAPM) says that the price of any asset is determined by the expected value of the asset itself, its covariance with the market portfolio, and the price (or cost) of risk reduction ($\Delta E(\tilde{V}_m)/\Delta\sigma^2(\tilde{V}_m)$). A firm – or any other asset provider – sells two things in the capital market: its expected future market value and the risk of that future market value.

7.4 Behavioral Uncertainty

Section 7.3 dealt with uncertainty emanating from nature. People can take actions to protect themselves from the consequences following from the occurrence of any particular state of nature, but they cannot affect the state of nature. It may be a good thing for people to be able affect the state of nature, but the analysis of such problems is more intricate because of the interactive character of the risks and actions. Risk emanating from the behavior of other economic agents has this interactive character, even in situations in which the true state of nature is not affected by behavior, only the outcomes facing some agents. The salient character of behavioral uncertainty derives from the fact that interacting agents have different sets of information and are either unwilling or unable to share it with their partners – or adversaries – in economic transactions. We offer a relatively long section entitled "Asymmetric Information," in which we discuss several distinct problems, and a relatively short section on strategic behavior, which would take us directly into game theory if it had not been decided not to embark fully into those wide and deep waters. Nevertheless, many of the asymmetric information problems, and correspondingly their solutions, have strategic aspects which formally have game theoretic structures.

7.4.1 *Asymmetric information: problems and solutions*

The provision of many goods and services involves situations in which the buyer and seller have different information, possibly about the product, about themselves, about the state of the world. The uncertainties that these informational asymmetries raise commonly prevent Pareto-efficient transactions from occurring. They also affect contract terms, quantities transacted, and prices and incomes. In many of these problems, the sequence of actions in the interactions between buyers and sellers is important. The timing of the arrival of information relative to the need to reach decisions is key to the informational uncertainty of these transactions: if I only knew whether you'll be a productive or unproductive employee before I hire you, I could make a better decision – I might not hire you! If you're unproductive and want a job, you don't want that to happen. If you are in fact productive, you'd want to tell me, but why should I believe you? You'd have the same incentive to tell me you are productive even if you aren't. The concept of the information set is the information relative to making a particular decision prior to (or at) the time of making the decision.

Many asymmetric information situations have an implicit temporal dimension that is not incorporated in static price theory, such as occupied much of the first five chapters. Since information is revealed over time (or over interactions, which themselves occur over time), the likelihood of repeated interactions among agents is important, as a source of information for buyers, as a source of potential profit for sellers. What an individual will do in a one-shot transaction may be quite different from what he would feel was prudent if he knew he was likely to repeat the transaction with the same agent many times – or with many agents if each of the buyers talks to the others about his transactional experiences.

I have divided the topics into broad categories of asymmetric-information structures that include both the problem, from a technical point of view, and the methods that agents in and out of markets have devised for conducting transactions despite the difficulties. I start with

the agency, or principal–agent, problem, which is quite extensive as a social phenomenon: one person – the principal – hires or commissions another to do a job for him or her. Considering the many advantages of division and specialization of labor, the principal–agent relationship is a common phenomenon. This relationship – it's not really correct to call it a "problem" all by itself – frequently is subject to moral hazard, which itself is a separate issue and definitely is a problem in the sense that it impedes efficient contracting. Moral hazard involves inability to observe the actions of a party to a contract. One classic moral hazard problem involves full-coverage insurance contracts: if your bicycle is fully insured against theft regardless of your own possible negligence in protecting it, you have no incentive to lock it up, literally encouraging thieves to steal it. Some writers equate the moral hazard problem with the principal–agent relationship, for example, Salanié (1997, Chapter 5) and Macho-Stadler and Pérez-Castrillo (1997, Chapter 3). The other major type of problem deriving from informational asymmetry is called adverse selection, which we discussed in passing in Chapter 3. Adverse selection occurs when a buyer cannot observe important characteristics of or contingencies affecting the seller, with the consequence that product quality may decline and some products even disappear from the market. Typical examples are difficulties in distinguishing between competent and incompetent doctors and lawyers and between good used cars and "lemons." Signaling and screening refer to different approaches to identifying one's own characteristics. The most prominent examples to date come from labor markets in which job seekers may want their productivity characteristics recognized so they can be hired and rewarded for them, while potential employers want to hire employees at wage contracts corresponding to their true productivity. How can truthful indicators be transmitted?

Throughout the exposition of these topics we will encounter a range of devices used as at least partial solutions to executing satisfactory transactions. These include the establishment of reputation, the use of contingent contracts, warranties, certifications, and monitoring. Always

remember that the importance of finding satisfactory transactions is that if only unsatisfactory ones are found, no transactions at all may occur and welfare will be reduced correspondingly.

The principal-agent relationship and the moral hazard problem

In a principal-agent relationship, the principal sets the rules and the agent follows them – more or less, which is the source of the problem when it occurs. Typically, a principal may be located at a different place from where she wants certain tasks accomplished, or she may lack the skills to execute the task as well as some other person who could do it better and at lower opportunity cost to the principal. (Many of us *could* accomplish tasks that we would prefer to hire out because someone else should be able to do them more efficiently, or because, even if we could do them more efficiently, we'd have to give up doing something else that we couldn't get done or that we'd just enjoy doing.) Some principal-agent relationships might be quite simple: the actions to be taken by the agent are observable, and it is straightforward to determine the outcome of the actions and specify the payment. The more problematic principal-agent relationships are characterized by the combination of poor (or non-) observability of the actions on the part of the agent and the joint influence of the agent's actions and stochastic events on the outcome of the actions. In this latter type of case, the agent could shirk – provide low effort when high effort was needed to ensure the desired outcome – but a state of nature could occur that would make it difficult to determine whether the poor outcome was the fault of the shirking agent or unfavorable nature. When sharing risks in a moral hazard situation, the private actions of the agent can affect the probability distribution of outcomes, but the contract must be signed before the agent has any incentive to take any action.

The primary informational asymmetry in the principal-agent relationship is that the principal cannot know either in advance or after the fact how much effort or care the agent expended in the contracted task. Since care or effort is so hard to observe, there will be temptations to under-supply it. Nonetheless, something other than the

effort or care expended must be unobservable or the quality of outcome could be equated with the care or effort. Frequently, controls are unavailable in advance, and observability after the fact is muddled by the simultaneous influences of the agent's actions and the stochastic state of nature. Quality of outcome may be debatable – hard to verify, making standards useless as a control device. Moral hazard frequently occurs in service provision, and services, once provided, can't be returned. Any information about either the actions taken or the state of nature, even imperfect information, can be used to improve principal-agent contracts involving moral hazard, accounting for the extensive detail sometimes found in contracts such as sharecropping agreements.

The key to the principal-agent relationship – how productive it will be – is the contract. When efforts are unobservable, a contract specifying them would be unenforceable and consequently useless. In the starkest case, the only observable phenomenon in the relationship is the outcome, so the relationship of the outcome to the agent's (and hence the principal's) payment, and the risk aversion of the two, are key. If the principal is risk neutral and the agent is risk averse, efficiency requires that the agent receive a fixed wage, independent of the result obtained but dependent on the characteristics of the task to be accomplished, the principal keeping the residual income. If the principal is risk averse and the agent is risk neutral, the best contract is a franchise, in which the agent pays the principal a fixed fee and keeps all the residual income. If both principal and agent are risk averse, the optimal contract will distribute the income risk between them according to their degrees of risk aversion.

To see clearly the importance of the asymmetry of information in the principal-agent contract, we develop the problem formally first for the case of fully observable efforts on the part of the agent. The principal contracts with an agent to execute a task, the object of which is the production of a particular outcome, x_i . It is a random variable, the occurrence of which can be affected by the agent's effort, e . The probability of outcome x_i occurring is $\text{Prob}[x = x_i | e] = p_i(e)$, which is read "the probability that event x takes the value x_i ,

given level of effort e ." The principal will pay the agent according to the outcome that occurs with a payment $w(x_i)$, which denotes that the payment w is contingent on the outcome x_i , ultimately on the agent's effort as well. Let the principal's utility be a function only of the income she receives from the outcome of the agent's effort, as directed, of course, by herself: $G(x_i - w(x_i))$, or, the outcome minus the payment to the agent. The agent's utility is a function of both his share of the outcome and his effort, the latter entering utility negatively. We could write the agent's utility function as $u(w, e)$; to simplify the analysis we will make the function separable in income and effort by specifying that $u(w, e) = u(w) - v(e)$, where $u' > 0$, $u'' < 0$ (positive, decreasing marginal utility of income) and $v' > 0$, $v'' \geq 0$ (marginal disutility of effort is positive, and things get no better, and possibly worse, as effort increases).

Now, the principal gets to set the contract, which is equivalent to saying that we shall set up the maximization problem from the principal's perspective. She maximizes her expected utility subject to the design of a payment contract with the agent that will give the agent at least what he could get in alternative employment: $\text{Max}_{e, w(x_i)} \sum_i p_i(e) G(x_i - w(x_i))$ subject to $\sum_i p_i(e) u(w(x_i)) - v(e) \geq u_0$, where u_0 is the agent's opportunity cost. This is called the "participation constraint," as it describes the minimum condition required to get an agent to participate in the venture. Set up the Lagrangean and take its first-order condition with respect to the wage contract $w(x_i)$, which gives the result that $\lambda = G' / u'(w) > 0$, where λ is the Lagrange multiplier on the agent's wage-payment constraint (the participation constraint). (We can infer the wage by inverting the marginal utility $u'(w) : w = (u')^{-1}$, where the exponent -1 indicates that we have inverted the function rather than taken "one over" the function. And remember that while u gets larger as w gets larger, u' , the marginal utility, gets smaller as w gets larger, because of the decreasing marginal utility implied by $u'' < 0$.) This result says that the ratio of the principal's and the agent's marginal utilities of income should be equal to a constant in the contract that optimally distributes risk between them; the constant is the marginal value to the principal of the utility the agent derives from the wage payment; if the agent derived more utility

from a given wage payment, the principal would need to pay him less to retain his services. This first-order condition is a constant characteristic of the optimal wage contract regardless what becomes of effort.

Now we compare this result to the principal-agent contract under the condition of moral hazard: since effort is not observable, it cannot be used as an optimization variable, and an additional constraint, called the "incentive compatibility constraint," must be used. When a contract is offered it is necessary to take into account the effort decision that the agent will make if he accepts the contract. The uncontrollability of effort decisions within the contract imparts an efficiency loss to the transaction, as well as affecting the type of contract that an agent will sign and the effort decisions she will make. The optimal contract now is a tradeoff between an efficiency objective on the part of the principal and the incentives of the agent. For the resulting contract to have any influence on the effort decisions of the agent, it has to pay him more when the outcome is favorable *and* when the favorable outcome is also a good indicator that his efforts were good or his decisions the correct ones. If a favorable outcome is only a weak guide to whether or not the agent made correct decisions or used a high level of effort – say, 80% of the time the outcome would have been favorable whether the agent worked diligently or slept – there is not a compelling reason to reward him especially highly under favorable outcomes. However, if the favorable outcome is distinctly – but not totally, or there would be no moral hazard problem – unlikely to have occurred unless the agent made good decisions/efforts, rewarding him correspondingly for the cogency of his efforts will be efficient; not rewarding him would lower the likelihood of securing his efforts in the first place.¹⁴

In this setting, the principal's objective function remains the same except for the control variables: $\text{Max}_{[e, \{w(x_i)\}]} \sum_i p_i G(x_i - w(x_i))$, where the square brackets surrounding e and w_i in the list of control variables indicates that the effort e is a control variable of the agent, not of the principal who controls only the variables within the braces – w_i . The participation constraint is the same as in the symmetric information principal-agent problem which posed no moral hazard. The

incentive compatibility constraint is actually a maximization problem of the agent, in which he chooses the level of effort that will maximize his own utility. As effort is not verifiable, the agent chooses his own level of effort, but to design a viable contract the principal must account for the agent's incentives. This constraint is that the value of e used in the objective function and the participation constraint be the level of e that maximizes the agent's utility: $\text{Max}_e \sum_i p_i(e)u(w(x_i)) - v(e)$. Thus anticipating the agent's behavior, the principal designs a contract that is the solution to this problem. Solutions of this general form are difficult, so to make the comparison of the symmetric and asymmetric information cases (or, without and with moral hazard), we can restrict the possible levels of effort (or range of agent's decisions) to just two – high effort e^H and low effort e^L , and corresponding outcome probabilities p_i^H and p_i^L . The principal will demand high effort e^H since any level of fixed payment would be sufficient to elicit low effort e^L . Consequently, the incentive compatibility constraint becomes $\sum_i p_i^H u(w_i) - v(e^H) \geq \sum_i p_i^H u(w_i) - v(e^L)$, or $\sum_i (p_i^H - p_i^L)u(w_i) \geq v(e^H) - v(e^L)$. The participation constraint is $\sum_i p_i^H u(w(x_i)) - v(e^H) \geq u_0$, because the principal wants to contract only for the high level of effort. The principal's objective function is $\text{Max}_{\{w(x_i)\}} \sum_i p_i^H(e)G[x_i - w(x_i)]$. Now, set up the Lagrangean for this optimization problem, using λ for the multiplier on the participation constraint as in the symmetric information problem and ϕ for the multiplier on the incentive compatibility constraint. The first-order condition with respect to the wage contract $w(x_i)$ can be written as $G'/u'(w) = \lambda + \phi[1 - (p_i^L/p_i^H)]$. We want to pay particularly close attention to the term multiplied by ϕ , which is the effect of having to consider the incentive compatibility constraint when designing the wage contract. If $\phi = 0$, $w(x_i)$ is a fixed wage, which elicits only e^L . The multiplier ϕ is the shadow price (cost) of the incentive compatibility constraint, and is strictly positive. The wage payment under moral hazard is higher, so the principal's profits were greater when the information on effort was symmetric than in a situation of moral hazard.

The ratio of outcome probabilities is worth remarking on. It indicates the precision with which the result x_i signals that the effort level was indeed e^H rather than e^L ; it is commonly called

the "likelihood ratio." A smaller likelihood ratio implies a higher probability that the effort was e^H when outcome x_i is observed. Working through the arithmetic, and remembering that u' gets smaller when w gets larger, a smaller likelihood ratio yields a higher G'/u' and consequently a larger wage.

Adverse selection

When the buyer is unable to observe characteristics of the seller or his product, we have the situation known as adverse selection, in which the quality of sellers and products gets self-selected out of the market, or if one prefers to consider nonmarket situations, the offers from higher quality suppliers dry up. In the case of either products (used equipment such as vehicles) or services, buyers are unable to distinguish quality differentials by any external characteristics. For any given proportion of "good" units and "lemons," a buyer will offer an average price for all units that adjusts for her expected likelihood of getting a lemon. This average price will drive the market price of all units of the product below what the suppliers of high-quality units are willing to accept. The good units of the service tend to stay out of the market. In the case of service providers, the expectation of obtaining the poor quality service may be high enough to keep more costly, but higher quality, service providers out of the market because they would be unable to distinguish themselves from the low-quality providers and hence unable to command a price sufficient to cover their costs. Adverse selection can keep the returns to a particular activity down low enough that it does not pay individuals to invest in training to become high-quality suppliers. Today we put people in jail for turning back the odometer readings on their automobiles before they sell them. In Classical Athens, law required the seller of a slave to make it clear if the slave were suffering from any sickness that would not be immediately apparent but could be recurrent and disabling; diseases that had to be declared included tuberculosis, strangury, and epilepsy (Westermann 1955, 16). In the Roman Empire, efforts were made to enforce the declaration of such debilitating diseases in slaves prior to sale (Westermann 1955, 99). These were efforts to reduce the lemons problem. Another practice in

Classical Greece encouraged it: while runaway slaves were punished severely when caught and returned, there was a reluctance to brand them as runaways: as Westermann (1955, 23) expressed it nearly poetically, “the practice [of branding recaptured runaways] was generally avoided in the Greek world because of the difficulty of selling a slave thus publicly advertised as having the tendency to disappear.” Descat (2011, 213) offers a brief, recent overview of the development of these laws and norms.

As with the case of moral hazard, contingent contracts may provide a mechanism for high- and low-quality suppliers to identify themselves. Warranties are one type of such a contingent contract. A supplier of a good who knows it will not last a week will not offer a two-month warranty. The supplier willing to offer the two-month warranty identifies himself as a supplier of the high-quality version of the good, while the supplier unwilling to offer such a warranty also identifies himself. Long-term agreements can be used as a quality-assuring device by giving suppliers a sufficient time horizon over which to recoup investments in higher quality. Lengthy contracts can also encourage slacking, however, so term limits themselves can control the supply of quality, from the bottom, so to speak.

Development of reputation is another mechanism that can identify quality differentials. Reputation consists of observable (fully known) characteristics of a seller’s past behavior. The seller has concern about the future demand for his services, which gives her incentive to exercise care in current transactions to provide a high-quality product. From the buyer’s perspective, reputation is a forecast of the seller’s future behavior – that is, her performance in the transaction the buyer is just now thinking of entering. Reputation earns a rent, so current stinting on effort would save a cost lower than the price obtainable by turning in a good effort. The size of the price premium will depend on the discount rate and the speed of information transmittal, which itself depends on the frequency of patronage by customers. Thus in relatively small markets, the rent may remain higher than in larger ones with a more thorough turnover of information. These rents imply a type

of inefficiency which appears as an excessively high cost for quality rather than overall inferior quality. Nonetheless, this higher cost of quality will force many consumers to purchase lower quality than they would be willing to buy if quality could be enforced contractually. With reputation as a signaling device, price can signal quality, at least partially – but generally only partially – repairing the damage of the asymmetric information that caused the adverse selection problem in the first place.

Signaling and screening

People typically possess private information about their own characteristics. The asymmetry of the information in this case derives from the twin facts that this information is not directly observable and that it may be possible – and profitable – to impart false information on those characteristics. For that reason, simple statements regarding the characteristics are not credible. Signaling and screening are two approaches to solving this particular asymmetric information problem. Signaling involves the individual with the private information sending a costly signal, the very costliness of which gives credibility to the signal. In signaling, the person with the private information sends the signal containing the information to persons who (may) want it, after which those persons make their offers (typically employment offers to workers who demonstrate different levels of competence or productivity). Screening reverses the order of moves between the person with the private information and the person who wants to know it.

The person wanting to know the characteristics of various people creates a menu of contracts that will let the individuals sort themselves (screen themselves) according to their unobservable characteristic. The following expositions of signaling and screening are adapted from those of Macho-Stradler and Pérez-Castrillo (1997, Chapter 5) and Molho (1997, Chapters 5 and 6).

Signaling

We deal with signaling first. The differential costs of different signals make the signaled information credible. The most common examples

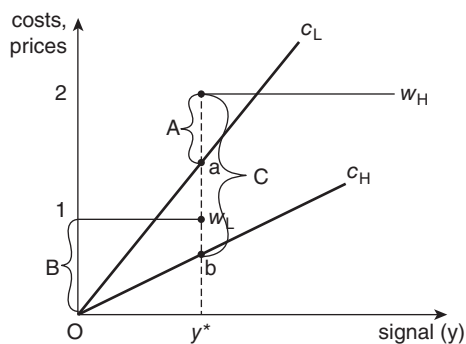


Figure 7.24 Signaling 1: separating equilibrium. Adapted from Molho 1997, 68, Figure 5.1. Reprinted with permission of John Wiley & Sons, Ltd.

of signaling problems come from the labor market – or any employment situation in which the employer initially does not have information regarding the competence or productivity of potential employees. Education is the signal that workers send. The costs of education differ in the sense that the more competent or productive workers are able to obtain the educational credential more cheaply than are the less productive workers, possibly because they are faster learners. Figure 7.24 shows how signaling works in the case of two types of worker. The vertical axis measures costs – costs of acquiring the signal on the part of the potential workers and the wage offers made to them by the employer. We measure the size of the signal, y , on the horizontal axis: either educated or not, $y = y^*$ or $y = 0$. The cost of producing the education signal for the high-productivity workers is $c_H = y/2$; the cost for the low-productivity workers is $c_L = y$. The rays c_H and c_L from the origin of the graph represent the costs of providing the signal y . The wages offered to high- and low-productivity workers are known beforehand as $w_H = 2$ for the former and $w_L = 1$ for the latter. The potential workers choose their education / signal levels, given the cost of education and the known wage differential. The low-productivity person *could* acquire the education required to obtain the high wage, but he is better off acquiring no education. We show this in the diagram.

The quantity of the signal represented by the education is indicated by the dashed vertical line y^* . The ray c_L cuts y^* at point a ; the distance from the horizontal axis to a is the low-productivity worker's cost of producing the educational signal y^* . The high-productivity worker's cost is the distance between point b and the horizontal axis along the vertical line from y^* . With zero cost invested in education, low-productivity workers can earn the wage equal to 1, yielding them distance B as a "profit" on their (non)investment. As we said, they *could* invest in education and receive the wage equal to 2, but at their cost of acquiring the education, they would receive only distance A as a profit. They received higher profit with no education and wage = 1 than with education and wage = 2. The high-productivity workers receive distance C as their profit, which is more than they would net if they were to forego acquiring education and take the wage = 1. Both groups of workers have no incentive to alter their signals in equilibrium, given the signaling costs and the wage offers. The employers make competitive wage offers and confirm their beliefs about the productivity levels each offer attracts. This particular type of equilibrium is called a separating equilibrium because the different types of workers are clearly separated in the solution, and it does not depend on the proportions of the different types of workers. These equilibria are not unique, however: the high-productivity workers could acquire more education than just y^* ; in this particular case, any amount between $y^* = 1$ (the minimum amount required to get the higher wage) and $y^* = 2$ (at which point the entire benefit of the wage is eaten up by education costs).

It was the relationship between education (signal-acquisition) costs and wage offers that permitted the two types of workers to sort themselves uniquely to educational levels and wages. With a different configuration of relative signal costs and benefits, the two types of workers could end up sending the same signal and getting the same wage offer, and we have what is called a pooling equilibrium, since both types of workers are pooled together in the outcome. Figure 7.25 shows the situation in which the wage

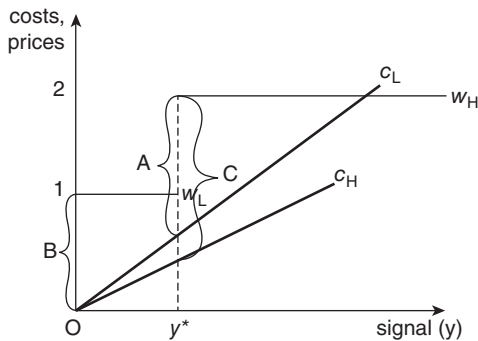


Figure 7.25 Signaling 1: pooling equilibrium. Adapted from Molho 1997, 73, Figure 5.2. Reprinted with permission of John Wiley & Sons, Ltd.

offer changes from 1 to 2 at a lower level of y^* than in Figure 7.24. Now the low-productivity workers find it more profitable to acquire the education signal (they get distance A from doing so) than foregoing it (distance $B < \text{distance } A$). The high-productivity workers still find it profitable to invest in the education, getting return C . Anticipating the prospective workers' behavior, the employer will form the following set of beliefs about these signals: if education $< y^*$, the worker is definitely low productivity; if education $\geq y^*$, there is a probability $= x$ that the worker is high productivity and a probability $1-x$ that the worker is low productivity. The higher wage offer will depend on the employer's probabilities, and the equilibrium does indeed depend on the proportions of high- and low-productivity workers. Multiple equilibria follow from the multiple beliefs employers might hold. Employers could develop interpretations of out-of-equilibrium signals that served to reduce the number of possible pooling equilibria. For instance, an employer could form beliefs about education levels $> y^*$ that would raise the probabilities she assigns to the high-productivity inference for higher levels of y^* .

Screening

The other approach to the problem of credible revelation of private information relies not on a signal sent by the person holding the information but on the particular contract the individual picks. The underlying problem is the same; the order of "moves" by the "players" in the "game"

differs. Again using employment as the example, employers offer a menu of contracts as combinations of wage and education pairs (w, y) . Workers select from the contract for which they qualify, the one that makes them best off. In equilibrium, no contracts are offered that will confer expected losses on employers, and no contract that could be profitable to an employer is not offered. The critical difference between screening and signaling is the order of the moves by the informed and uninformed parties. No pooling equilibria exist under screening. Separating equilibria would be characterized by sets of points (pairs of (w, y) contracts in the two-productivity-type case) in the cost / price-education / signal space of Figures 7.24 and 7.25.

As to which model is "correct," different markets might find it advantageous for one or the other of the informed and uninformed agents to make the first move. Insurance markets seem to fit the screening model – companies offer contracts and customers choose. In the case of employment, it may be more reasonable for individuals to choose signals prior to seeking employers.

Conison (2012, 100–121) uses the screening concept to analyze how Roman law developed to deal with informational asymmetries among sellers and buyers of wine. He develops a game-theoretic model to reveal that, in the incomplete contracts prevalent, the combination of the *degustatio* (tasting) with a default rule assuming that tasting had been conducted provided a separating equilibrium which helped buyers distinguish, if still imperfectly, a seller's attempt to cheat them on wine quality.

7.4.2 Strategic behavior

As we said earlier, the uncertainties of an agent's interactions with other agents are the most intricate sort of risk that economic agents face. As disastrous as natural risks can threaten to be, and as expensive as some of the *ex ante* and *ex post* protective actions may be, they are relatively straightforward. When economic agents interact strategically with other economic agents, each agent may recognize that not only are his actions subject to prediction by other agents but his own predictions of their predictions about him enter into his calculations supporting his decisions.

The asymmetric information situations discussed in the first part of this section have elements of strategy surrounding the decisions the actors make, although we did not emphasize those aspects of the problems. We noted the timing of when various agents knew particular pieces of information, and the order of actions taken by the informed and uninformed agents, as well as the potential or actual importance of repeating the interaction in both the principal-agent problem and in reputation formation. These phenomena are the substance and subject of game theory, which despite its frivolous sounding name is the serious, mathematical analysis of strategic interdependence. It has been used in political science, anthropology, ecology, and in recent years extensively in economics. This subsection will introduce some of the basic concepts of game theory as applied to economic interactions, but because game theory very rapidly gets demanding mathematically we will just skirt around the shores of that sea: the “shelf” is narrow, the “drop off” is abrupt, and the “bottom” is deep.

There are three major elements of a game.¹⁵ First, there are the players – the agents interacting in the game. Second are the rules: who “moves” (acts) when, their knowledge set when they move, and the actions that are open to them. Last are the outcomes or payoffs: the results of each possible set of actions and the players’ preferences regarding (utility from) those outcomes. These components are sufficient to define quite intricate problems of interdependent choices. Agents face uncertainties about the actions their rivals will take and consequently uncertain outcomes for themselves. Nature itself may be considered a player, but only to add another, noninteractive set of uncertainties.

It will be useful to identify some categorizations of games, even if we do not delve into the solution methods of them. These typologies themselves should heighten readers’ awareness of the types of situation they may face in trying to step into the shoes of ancient decision makers. The simplest type of game is the static, full-information, simultaneous-move game. All the players know who all the other players are and the payoffs of each player in each possible outcome. Depending on the pattern of payoffs, every player’s decisions will be completely clear – deterministic; these are called “pure” strategies. For some configurations

of payoffs, some players’ choices will not be obvious, so all the players will have to assign probabilities to those players’ actions in order to assess their own best strategies; these are called “mixed” strategies. Having used the word “strategy” in a context that comes close to being a technical one, we offer a definition of the term in the game-theory context: a complete, contingent plan or decision rules specifying a player’s action in every possible, distinguishable circumstance throughout the game. The role of the qualifier “distinguishable” may not be clear yet, but we will make it so shortly.

Static games are represented in what is called “strategic” form, or “normal” form – a tabular form, as shown in Table 7.3. The players are identified as players 1 and 2. Player 1 has actions A and B available to her, and player 2 has actions X and Y available to him. We could give any set of names or symbols to the players and any set of names or symbols to their possible actions. The four cells in the lower right corner of this table represent the payoffs to player 1 and player 2, respectively, if they should choose the set of actions corresponding to that cell. Thus, if player 1 chooses action A and player 2 chooses action X, both players get a payoff of 1. If player 1 chooses action A and player 2 chooses action Y, player 1 gets a payoff of –1 and player 2 gets a payoff of 2 (a better outcome for player 2 than the combination AX, a worse outcome for player 1). The concept of a “solution” to this game, or to any game, is to find an equilibrium set of actions, equilibrium in the sense that no player would choose a different set of actions when faced with the same set of possible outcomes.

The game in Table 7.3 is the famous “Prisoner’s Dilemma” game, which it will be useful to discuss. In this game, the two players have been arrested and imprisoned for a crime, but the arresting authorities lack the evidence to convict either one. Consequently, the police, as

Table 7.3 A payoff matrix.

		Player 2	
		X	Y
Player 1	A	1, 1	–1, 2
	B	2, –1	0, 0

we'll call the "arresting authorities," interrogate each prisoner separately to attempt to get each to give evidence against the other. They do not let the prisoners communicate with each other for the obvious reason that if they did the suspects would get to learn what each other's stories are. In the technical lexicon, their information sets would change from no information to full information, while that of the police would remain at no information. The police tell each prisoner that if he cooperates by ratting on the other, he will be freed and rewarded for his testimony. If one testifies and the other does not, the one who testifies goes free with some change in his pocket and the other stays in prison, but if they both testify, they both remain in prison but they do get the cash reward for testifying. (This may seem like a strange deal, feeling like you got a positive payoff when you stay in prison, but remember that these payoffs are simply the results of putting the outcomes – cash, prison time or both – into a utility function. We could assign different numbers so as to yield zero or negative evaluations of prison time, but as long as the configuration of payoffs is as in Table 7.3, the incentives are the same.) However, if neither testifies, they both will go free for lack of evidence, but without the money in their pockets. What do they do? Experimentation with this game indicates that when it is played once, both players tend to confess; they both get the cash payment but they stay in prison. The reason can be seen clearly by looking at the alternative facing each player. If player 1 were to try to maintain solidarity with his crime mate by keeping his mouth shut – action A – he stands to get a payoff of 1, but he knows that player 2 stands to get a payoff of 2 by confessing – *and* player 1 knows that if player 2 takes that opportunity to improve his wellbeing, he will get a payoff of negative 1 rather than positive 1 – not only does he stay in jail, he forfeits the cash payment. So the combinations of AX and AY are ruled out by player 1. Player 2 faces a symmetric set of rewards and penalties, and he rules out strategies AX and BX. This leaves strategy BY, which imprisons them both but with the reward, the disutility of imprisonment canceling out the utility of the cash payment. Each could have received a

better deal without harming the other had each trusted the other to not be greedy and try to take advantage of the goodwill extended by the choice of strategy AX.

The name "Prisoner's Dilemma" derives from just one possible story that sets up this pattern of payoffs to a set of strategy options, but many possible economic interactions fit this payoff pattern. In an important extension of the game, known as a repeated game, the inefficiency of the one-shot Prisoner's Dilemma game – both players leaving available payoff on the table – can be remedied. Repeated games are important in a number of the asymmetric information problems we discussed in the previous subsection. In the Prisoner's Dilemma, repetitions give the players the opportunity to send signals about their willingness to cooperate and about the penalties they are willing to confer on the other if the other deviates from the equilibrium that yields the jointly highest payoff. Thus in the Prisoner's Dilemma, over a number of repetitions, the pair of players can learn that they can both be better off if they both keep their mouths shut. But, in the repeated-game context, if one of them confesses and the other doesn't in one play of the game, the other can retaliate the next time. After a few plays of the game, the players get the hang of the game and keep their mouths shut and go free without the reward. It would be unfortunate if readers took away the message that we are talking only about ways to get out of prison. The game structure is much more general, representing opportunities for economic agents to gain trust in one another and learning not to take short-term advantage of each other if for no other reason than that the other can retaliate at the very next opportunity.

The other major category of games is the dynamic game, with or without full information. Dynamic games may include simultaneous moves, but they also involve a sequence of interactions that can take place over time. Common problems studied in economics with dynamic games are the issues of entry of new sellers (firms) and deterrence of entry by incumbent sellers (firms), common problems in duopoly and oligopoly settings. Strategies (actions A and B

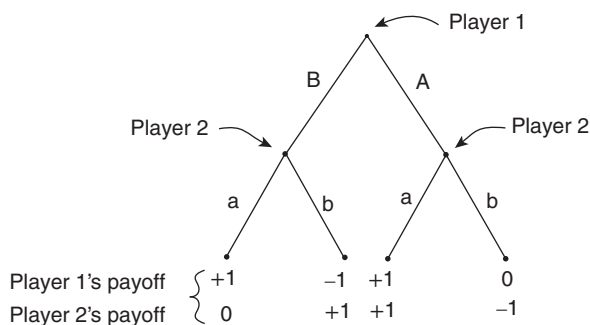


Figure 7.26 The extensive-form representation of a game.

and X and Y) correspond to choices of output and pricing. Correspondences to political problems are obvious: deterrence of attacks by outsiders, when to attack and how, threats between states and the establishment of specific terms in treaties. Dynamic games are represented in what is called “extensive” form, which we show in Figure 7.26. The diagrams are combinations of nodes, representing decision points for specific players, and lines representing the decisions. The extensive form of a game has a dendritic appearance. In Figure 7.26, player 1 makes the first move, choosing either action A or action B. Then it is player 2’s turn, with the choices a and b for either of player 1’s actions. This particular game is over after player 2 chooses either a or b, and the payoffs for each player are noted below the bottom nodes. Thus, if the sequence of actions is Ba (action B by player 1 and action a by player 2), player 1 gets a payoff of +1 and player 2 gets a payoff of zero. Conison (2012, 118 Figure 3.1) employs such an extensive form game graph in his analysis of the Roman wine industry’s degustation.

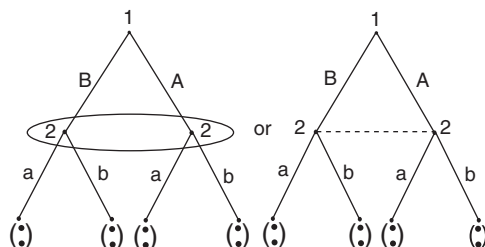


Figure 7.27 Extensive-form representation of a game with limited information.

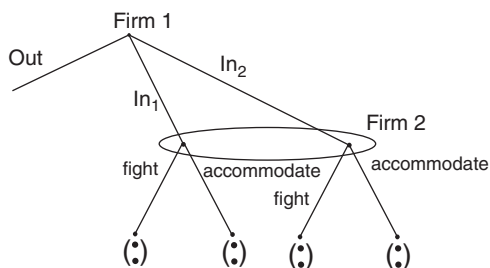


Figure 7.28 Representation of a firm's entry decision with incomplete information.

Figure 7.27 shows a game with the same structure of moves but shows the two alternative ways of indicating that when it comes time for player 2 to make his choice of actions, he does not know whether player 1 chose A or B; thus player 2 doesn't know if he is at the left-hand or right-hand node. This incomplete information can drastically change the character of the choices in the game. Figure 7.28 shows another game structure, characterizing the problem of a firm deciding whether or not to enter a new line of production. If Firm 1, the potential challenger,

decides not to enter this line of production, it takes the left-hand leg from the first node, ending the game. However, if it chooses to enter, it has two methods of attempting to make the entry, identified as “In₁” and “In₂,” which could refer to the level of production at which it plans to make its entry attempt. Once Firm 1 has made one of the “In” decisions, Firm 2 cannot distinguish immediately which entry plan Firm 1 has made, only that it has decided to challenge its place

in the production of this good. Firm 2 has two options for its response to Firm 1's entry: it can accommodate Firm 1 by retrenching a bit on its output or by fighting Firm 1's entry by cutting its price. We have not assigned payoffs to the four strategies, but they certainly would differ. Firm 2 may have to make a decision that will affect its profitability or even its survival on the basis of incomplete information.

An issue that arises in dynamic games that did not exist in static games is that of credibility of a player's strategy. Some strategies would never be played, regardless of what other players might do along those decision paths. Thus, some threats – which are themselves proposed strategies – may not be credible.

The solutions of games correspond to profit maximization or utility maximization in the types of problems we studied in the first five chapters. The equilibrium includes a strategy for obtaining the equilibrium payoff, whereas the equilibria we discussed earlier consisted just of the payoffs. There are, additionally, strategies for determining equilibria in game-theory models. The simplest is the dominance strategy, in either a strict or weak form. The player looks for a strategy that always gives the best outcome, regardless of the actions that other players might choose. If one strategy always gives a superior outcome, we have strict dominance (one strategy strictly dominates others, in the sense that it gives higher payoffs). If the best one can do is find strategies that are at least as good as others but not necessarily better, we characterize the solution strategy as one of weak dominance. As we noted above, pure strategies can make these choices with certainty, while some payoff-action configurations may require players to characterize the probability distributions of payoffs, depending on the actions that other players might take, yielding the mixed strategy. Another important, alternative solution approach has players use what is called the Nash equilibrium: each player follows a strategy that is the optimal response to other players' best strategies.

Beyond this brief introduction to concepts, game theory gets deep quickly, and the technical details of solution methodologies commonly absorb as much of the attention as the substance of the problem under study. We believe, however, that an awareness of the strategic and

informational issues that game theory formalizes will serve readers well.

7.5 Expectations

We open this section with a brief exposition of why expectations are important in the study of economic decisions. The following two subsections develop an older and a newer approach to the formation of expectations. But first, what is an "expectation"? In the current period, economic agents frequently have cause to think ahead to the future when making the calculations involved in a decision. Variables whose realizations, in the probabilistic sense, have already been realized have known values: people can look around and see the current prices of goods, services, their own labor, and so forth. Those values are important in current decisions, but many forward-looking decisions are made on the implicit or explicit belief (expectation, in the probabilistic sense) that some of those prices (or quantities) will take particular values or ranges of values in the decision-relevant future. Quite specifically, the current price of, say, dates is x shekels per bag. We have the opportunity to plant some new seedling date trees that won't produce dates for a few years. In deciding how many to plant, we consider what we think the price of dates will be at the time our seedlings' dates appear. Those prices are current expectations of future date prices. If they're particularly low (maybe the king has just planted 1000 hectares in date trees), we may decide to plant none at all or only a few. If there has just been a flood that wiped out a large proportion of the region's date trees, you may decide that dates will have a good price for a few years, including at least the first few years that your new seedlings will produce fruit; after that, you suspect that everybody else's plantings will be maturing too, and there will be lots of dates, so the price will decline thereafter.

This example includes a number of factors that are important in theories about expectations formation. We will review an older approach to expectations formation – adaptive expectations – that left many components of the theory implicit and consequently gave predictions that didn't square well with empirical observations. This older model is well worth

studying, not the least as a bad example. It is unnecessary to throw rocks though, because, as is characteristic of the progress of any science, subsequent theorization built on a combination of its foundations and its observable shortcomings. The current approach to expectations formation is called rational expectations, which simply hypothesizes that economic agents use the best available information when making their forecasts. This said, we turn to how expectations are used.

7.5.1 *The role of expectations in resource-allocation decisions*

While many of the resource-allocation decisions discussed in the early chapters had most of their benefits as well as their costs concentrated in the current period – that is, the actual decision period – we began to explore decisions involving longer terms. Consumer durables are the leading edge of a large number of goods we could characterize as capital goods, even including labor, which has a number of capital-like qualities that we will discuss in a later chapter. The holding of inventories, ranging from stocks of consumer goods to a business's stocks of inputs and finished products, to money holdings, clearly involve expectations related to those goods. When considering current allocation options for resources that are expected to last into the future, decision makers make forecasts (by any other name, still forecasts) of variables that would make their current decisions more or less beneficial in the future. Imperfect and costly information gives rise to uncertainties about the future, and decision makers find it essential to form views about the future to make cogent intertemporal decisions. Current decisions depend on current evaluations of the future.

In the discussion in section 7.2.2 about current beliefs and Bayesian updating of beliefs, we were talking about how decision makers use new information. New information was filtered through a likelihood function, which itself yielded new information that modified prior beliefs. In this section, we develop some theories of how individuals endogenously form their expectations. These theories have the structure of a rule or set of rules by which agents revise their expectations in light of new information. This sounds very much

like the economics of the likelihood function. In another sense, theories of expectations are more restrictive than likelihood functions in their most general sense, because likelihood functions yield predictions about a much larger array of variables than do models of expectations.

The variables about which expectations models yield predictions are the endogenous variables in economic agents' implicit or explicit models of how components of their economic systems work. To get more specific, in the example above of planting date seedlings, our farmer was thinking about the behavior of the date market (or if you don't want to think in terms of markets, how the value of dates would behave in the future). Long- and short-term weather trends clearly are important, but these are strictly exogenous variables for the problem the farmer is trying to solve. Economic theory – the agent's economic theory, as well as the contemporary student's – isn't used to predict exogenous variables; the farmer does, of course, form expectations about the weather, but he doesn't use his economic models to make that forecast. His problem is to choose how many date seedlings to plant, and the model he uses to think about that problem will contain the influences he believes will affect the value of his dates in the future. This may seem like the beginnings of a controversial – or outright rejected – concept for ancient economic behavior: that ancient decision makers understood how their economies worked and made their economic plans on the basis of those models. We think this idea should not be controversial, although being explicit about it may prompt some new targets of research on those periods. In the first place, the archaeological record yields clear evidence of causal thinking in engineering statics/mechanics problems; it is unreasonable to believe that the same thought capacity did not extend to issues of social interactions involving livelihoods. Second, the written record yields ample evidence of mental models of the exogenous and endogenous influences on the values (magnitudes) of economic variables.

7.5.2 *Adaptive models of expectations*

The simplest assumption about what future prices will be is that next period they'll be the same as this period. We develop some notation to convey this relationship: $E_t(p_{t+1}) = p_t$. The

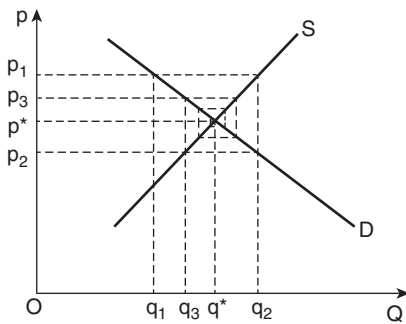


Figure 7.29 Expectations: the cobweb model of agricultural supply, convergent conditions.

$E(\cdot)$ notation is the expectation operator and is read “the expectation of.” The subscript t on the expectation operator indicates that the expectation is formed and held in time t . So the entire expression is read as, “the expectation at time t of the price at time t -plus-one is the price at time t .” The famous cobweb model of agricultural supply, which we show in Figure 7.29, uses this type of expectation model. Initially, supply and demand for this agricultural commodity are in equilibrium at price p^* and quantity q^* . Farmers expect price p^* to prevail in the next period and so plan to produce quantity q^* , but the weather is poor for crops and they manage to produce only q_1 . With this supply shortfall, the price that clears the market is p_1 . The farmers expect price p_1 to prevail thenceforward and, for the next period, plan to produce quantity q_2 . The weather holds, and they succeed in doing so, and of course the price falls to p_2 . Now, they figure that price p_2 will prevail in the following period, so they plan to produce quantity q_3 , which they also succeed in doing; the price for this quantity is, naturally, p_3 because the demand curve does not shift. You’ll notice, however, that these prices and quantities are approaching the old equilibrium. This model gives some rationale for a one-time disturbance from equilibrium to gradually, but not immediately, return to the initial equilibrium, in an oscillatory fashion – that is, the prices alternately overshooting and undershooting the equilibrium price. One problem with the model is that if the supply curve is flatter than the demand curve (supply is more

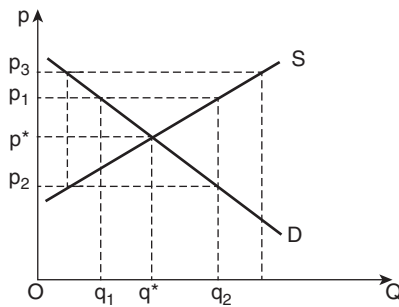


Figure 7.30 Expectations: the cobweb model of agricultural supply, explosive conditions.

responsive to price than is demand, a characteristic of agricultural products), the displacement from the initial equilibrium will be explosive, as Figure 7.30 shows. After the initial displacement to the out-of-equilibrium (p_1, q_1) , subsequent price-quantity pairs get farther away from the initial equilibrium. This is a bad characteristic of a model because the markets modeled don’t behave that way. There’s nothing wrong with the specification of the demand curve, but the supply curve has farmers plan their production decisions on the basis of quite naive expectations of future prices. Farmers aren’t stupid, and few people would suspect them of expecting the high prices associated with a severe drought – or excess rain – to prevail under normal weather conditions.

A somewhat more sophisticated model of expectations is that the current expectation of the future price is proportional to the difference between the current actual price and the previous period’s expectation for the current-period price: $E_t(p_{t+1}) = E_{t-1}(p_t) + \beta[p_t - E_{t-1}(p_t)]$, where the value of β is between zero and one. This is called the adaptive expectations, or error-learning, model. The change in price expectation is a constant percentage of the gap between last period’s actual price and the price that was anticipated for that period, or a weighted average of last period’s actual and expected prices. If β is small enough, the price will not oscillate around the equilibrium after a displacement but will approach it smoothly; but if $\beta = 1$, the adaptive expectations model becomes the old cobweb model, and we

still have the potential for explosive behavior after displacement from equilibrium. We emphasize that this explosive behavior predicted by these models in these circumstances doesn't really alert us to a particularly interesting social possibility, just a faulty model.

The regressive expectations model is a special case of the adaptive expectations model that makes the current expectation of the future price a weighted average of the price in a large number of previous periods, or equivalently, of differences between prices in adjacent periods. This can be written as $E_t(p_{t+1}) = p_t + \theta(p_t - p_{t-1})$, or $E_t(p_{t+1}) = (1 + \theta)p_t - \theta p_{t-1}$. If $\theta = 0$, $E_t(p_{t+1}) = p_t$, or the naive extrapolation of the current price. If $\theta = -1$, $E_t(p_{t+1}) = p_{t-1}$. Generally, if $\theta > 0$, past changes in prices are extrapolated and expected to continue, whereas if $\theta < 0$, past trends are expected to be reversed. This specification is a special case of $E_t(p_{t+1}) = \beta_0 p_t + \beta_1 p_{t-1} + \beta_2 p_{t-2} + \dots$, in which $\sum_i \beta_i = 1$ is a common restriction. An alternative formulation for the β_i lagged-price coefficients is $\beta_i = (1 - \lambda)\lambda^i$, with the value of λ being chosen to be between zero and one. This formulation of the effect of previous prices on expected future prices is known alternatively as the geometric distributed lag or the Koyk lag: $E_t(p_{t+1}) = (1 - \lambda)p_t + (1 - \lambda)\lambda p_{t-1} + (1 - \lambda)\lambda^2 p_{t-2} + \dots$. This last expression can be rearranged to yield an expression for how expectations change over time as a function of the difference between previous actual and expected prices: $E_t(p_{t+1}) - E_{t-1}(p_t) = (1 - \lambda)[p_t - E_{t-1}(p_t)]$, which is equivalent to adaptive expectations.

This has been a lot of complicated rearrangement of terms and intricate notation through which I have dragged you. The point of the trip through the previous paragraph is that these expectations models all use only past prices as their information base. The closest any of it comes to including information directly about the future is the parameter θ , which could take positive or negative values, but only arbitrarily according to whether either the agent or the student modeling the agent thought a trend was going to continue or be reversed, but without any explanation for that belief. During either an upward or downward trend, the agent makes systematic mistakes in his

predictions: either way, his expectations never quite catch up with (sometimes even get close to) the prices that actually occur. The structure of the expectations formation method prevents accurate predictions, even on average. This is the problem that the rational expectations model set out to correct.

7.5.3 *The rational expectations hypothesis*

The rational expectations (RE) model derives its name from the assumption that agents use the information available to them as efficiently as they can to make predictions about future values of relevant economic variables. All other economic models specify optimizing behavior; the RE model just applies this methodology to the formation of economic expectations. An implication of the RE model is that people don't make systematic mistakes in their forecasts. If they did, someone (say, a speculator) could benefit from taking advantage of that information; in doing so, she would bring the average realization into line with her expectation.¹⁶ In the previous class of expectations models we introduced, the agents took no action to revise their rule for forming expectations even though they made predictable errors.

According to the RE model, an agent's informed forecasts of future events (values of economic variables) are the same as the predictions of the relevant economic theory. An agent's subjective expectation, formed at time $t-1$ of the value of some variable X at time t , ${}_{t-1}X_t$, is the conditional expectation of that variable given the information available to the agent at time $t-1$, $E(X_t | I_{t-1})$. This specification of the expectation formation process lets the agent break out of the past-values-only straitjacket. If our date-planting farmer's information set includes the knowledge that everyone in the kingdom is planting even more date seedlings than he is, his implicit model of the supply-demand process will give him the expectation of low date prices in the future. If his information set tells him that there are only four hectares in the kingdom suitable for growing dates and he has two of them, filtering that information through his implicit

supply-demand understanding will yield the information that prices look good from his perspective. An implication of rational expectations is that expectations about variables will change when the conditional probability distributions of those variables change. If our farmer starts with the latter expectation and receives the information that the king has signed a trade treaty with a date-rich, neighboring kingdom, he will revise downward his expectations of the future price of dates in his own kingdom.

At this point, it is useful to write out the formulations for the three expectational models we've discussed so far so we can compare them directly. In the cobweb model (which is the adaptive expectations model with $\beta = 0$, we can write the quantity demanded at time t as $Q_t^d = a - bp_t$ (as price p_t goes up, the quantity demanded goes down, the relationship we expect) and the quantity supplied as $Q_t^s = c + dp_{t-1}$, which says that the farmer makes his time- t production decisions on the basis of last period's price. We set the quantity demanded equal to the quantity supplied and solve for the equilibrium price in the current period as $p_t = (a-c)/b - (d/b)p_{t-1}$; solving that for the current price in terms of only parameters, we get $p_t = (a-c)/(b+d) - (d/b)^t$. If the absolute value of $-d/b$ is less than one, a disturbance from the initial equilibrium will return eventually to that price-quantity configuration. The adaptive expectations model explicitly introduces an expectation model, so we get an additional relationship in that model. The demand function is the same: $Q_t^d = a - bp_t$. The supply function looks pretty much the same but is actually complicated by the expectational form of the current-period price: $Q_t^s = c + d_e p_t$, where we use the notation $_e p_t$ to represent the expected price at time t (we drop the notation indicating that the expectation of the price at time t was formed at time $t-1$, which is implicit). Now, with a price-expectation variable in the supply relationship rather than an observed price, we need to include the relationship between actual and expected prices. The adaptive expectation relationship is, using the same, abbreviated notation for expected prices, $_e p_t = \beta p_{t-1} + (1-\beta)_e p_{t-1}$. We substitute this expression for the expected price in the

current period back into the supply relationship, equate quantity demanded with quantity supplied, and with some further rearrangements, get the following expression for the current, actual price: $p_t = (a-c)/(b+d) + [1-\beta(1+d/b)]^t$. As you can see, if $\beta = 1$, this is exactly the same expression as the cobweb model. Nothing of signal importance has changed between the two models.

Now, let's explore the rational expectations model of supply and demand. The demand equation is still the same: $Q_t^d = a - bp_t$. In the supply equation, we include the weather during the production period as well as the expected price: $Q_t^s = c + d_e p_t + kw_t$. The relationship for the price expectation is $_e p_t = E(p_t | I_{t-1})$. The information in the conditional expectation of the price is lagged one period because we do not yet have information from the current period (period t). Once again, we equate the expected quantity supplied with the expected quantity demanded, but beyond this point, we proceed somewhat differently in this case. Note first that there are no lagged prices in this model. Equating expected supply with expected demand gives us $a - bp_t = c + d_e p_t + kw_t$. We take the expectation of that relationship to obtain the following expression for the expected price in period t : $_e p_t = (a-c)/(b+d) - [k/(b+d)]_e w_t$. Note that the expectation of the weather (the forecast of what the weather will be during the growing season) is part of the agent's economic model. Now, we substitute this expression for the expected price in the current period (an expectation formed in the previous period without knowledge of the realized values of any variables from period t) into the formulation for supply: $Q_t^s = c + [d(a-c)/(b+d)] - [dk/(b+d)]_e w_t + kw_t$. Note that both the expected weather and the actual weather enter into the supply formulation. Now we proceed as before and equate supply and demand (the previous equation of supply and demand was what the agent did in his formulation of his expectation, knowing – from his implicit model of how agricultural prices “work” – that whatever price emerges will be a consequence of the equalization of supply and demand). The resulting, current-period price is $p_t = [(a-c)/b] \cdot \{1 - [d/(b+d)]\} - [dk/b(b+d)]_e w_t - (k/b)w_t$.

The resulting price is independent of previous prices; although we solved the previous models for the current price in terms of only parameters, some of those parameters (specifically the parameter d in both the cobweb and the adaptive expectations models) referred to the previous, realized price. As you can see by tracing the more intricate interaction pattern of the coefficients from the supply and demand relationships in the RE solution for the current, realized price, this independence from previous realized values is obtained by putting the supply-demand coefficients into the solution twice – once for the expectation and once for the realization.

This independence of the solution for the current, realized price (or any other current, realized variable) is not a necessary or even a general characteristic of RE models. In fact, a slightly more complicated agricultural supply-demand model with the possibility of storage of the product will contain the lagged realized price in the solution for the current price, but because of the inventory holding relationship, not because the price expectation mechanism is an extrapolation of previous price realizations.¹⁷

If we combine the concept of the forward-looking expectation with that of the covariance risk, we get a somewhat different perspective on farmers' attitudes to weather and pest risk. After living and farming in a location for several generations, farmers will have obtained and passed on considerable knowledge of the probability distributions of weather and pest "events," although farmers of two to four millennia ago (and even more recently) would not express their knowledge in such language. Certainly they will know the expected values (the means). Undoubtedly they will know the variances as well. A random occurrence of a weather or pest variable within so many standard deviations (the square root of the variance) of the mean will be recognized as not especially out of the ordinary. It will not be a "surprise." This does not mean that ancient farmers were statisticians, just that their experience would have been able to tell them – collectively – that certain events were likely about once every 20 years, although they equally certainly would hope that this year (and each year in its turn) wasn't the year. The expectation of such events will be

incorporated into their cropping plans. Evidence from nineteenth-century C.E. American farmers indicates that they were able to incorporate variability as high as the second moment of a probability distribution into their expectations; third moments were problematic (McGuire 1980, 347–349). Just looking at variances of crop yields or rainfall or pest incursions (from contemporary or recent data; ancient evidence is much sparser) will not tell us much about the farming risks that farmers faced. They may have expected much of it and adjusted their operations accordingly to smooth out their overall risk.

Additional characteristics of the distribution of weather and pest events are important in agriculture. Third moments measure the skewness or lopsidedness of a distribution. A symmetric distribution would have a third moment of zero. The most likely observations from an asymmetric distribution will be to one side or the other of those from a symmetric distribution. For example, while the extremes of rainfall might be from 44 to 144 mm per year, the most likely rainfall in any particular year with a symmetric distribution would be 94 mm, exactly midway between the extremes. The most likely observation in a skewed distribution of rainfall with the same minimum and maximum would be either below or above 94 mm per year. Fourth moments measure kurtosis, or the shape of the center of mass of a distribution, relative to that of a normal distribution. For example, the central tendency could be thin and needle-like or relatively flattened and dumpy rather than having the bell shape of a normal distribution. Depending on the distribution of the observations, such a distribution could be accompanied by either thin or fat tails, that is, the extreme observations could have either higher or lower probability of occurrence than those in a normal distribution. Beyond these patterns of likelihood, weather events over time may not be independent, but instead be autocorrelated, such that a high observation in one year would tend to be followed by a high observation in the following year (positive autocorrelation) or by a lower observation (negative autocorrelation). At least some of these characteristics of meteorological phenomena would have been in the information sets of ancient agriculturalists.

If some of those properties changed, farmers would have had the problem of assessing whether the changes were temporary or permanent. If the changes were permanent and farmers interpreted them initially as temporary, they could have made grave mistakes in their reactions.

The rational expectations hypothesis – it is a hypothesis, producing refutable predictions – does not imply that agents know what future prices are going to be, or that they are always right – only that they are right on average, otherwise they'd revise the procedure by which they formed their expectations. Neither does it imply that they use the correct model of the situation they are forecasting, although this statement gets a bit trickier. If their model is incorrect, their forecasts are likely to be systematically wrong – if the agents are able to figure out what part of the forecast errors are systematic, which involves some degree of analytical capability, they could figure out how to begin changing their forecast model. Clearly, highly sophisticated, time-series statistical methods which have been developed only in the past several decades were not available to ancient decision makers in the Mediterranean and Aegean region, regardless of their sophistication in engineering mechanics. Applications of RE to predictions of variables within a banking system – which we have not considered yet – are a case in which individuals and entire societies as recently as this century have operated – belabored we might say – under faulty models of “how the system works.” Nonetheless, the veritable industry that has been working on banking theory for the past century must stand as evidence that agents are working to improve the models that inform their expectations formation.

The use of the RE model of expectations formation for, say, Mesopotamian farmers of four thousand years ago neither means nor requires that they possessed concepts like demand and supply elasticities or demand and supply functions. What it does mean is that, before making a costly decision, those farmers – and shopkeepers, potters, donkey drivers, and others – said to themselves something like, “What’s going to happen if I increase my supply of X (whatever it is that I provide)? Is everybody else who provides that likely to do the same thing? If they do, my own advantage from increased supply is lessened according to how much the wants for what we

provide increase at the same time.” The level of consciousness of this type of calculation is an open question.

7.6 Competitive Behavior under Uncertainty

Uncertainty can be – and has been – introduced into a wide array of models that we have studied so far in the certainty case: production decisions; the demand for, and supply of, factors of production; consumption and saving decisions; interactions among firms; inventory holdings; and others. In closing this chapter on risk and information, we offer brief overviews of two particular topics – production decisions of the firm and search behavior.

7.6.1 Production behavior

Let’s think of our treatment of the competitive firm from Chapter 2, where the firm maximized profit as the difference between the value of its output and the cost of its inputs. There are three principal avenues for uncertainty to enter this problem: the price of the firm’s output, the price of its inputs, and the success of its productive process. A variant on the input price uncertainty is the possibility that the quantities of inputs have a random component not under the control of the firm. These three different sources of risk affect a firm differently, as the expected output price, expected input prices, and expected input-output relationships are located at different places in the Lagrangean used in the optimization problem. Additionally, specifying the firm’s problem as a profit-maximization problem implies risk-neutrality on the part of the firm, which may not characterize the particular firms a student wants to analyze. It is necessary, at the least, to set the problem up as a utility maximization problem, where utility is a function of the firm’s profit. However, the firm may pursue alternative objectives altogether, such as minimizing the likelihood of losses; maximizing the likelihood that profit is at least a certain level; or maximizing expected profit, but subject to a constraint on the likelihood that profit falls below a certain level.

With output price uncertainty, and all input and output decisions made *ex ante* (that is, before the

uncertain prices are known), a risk-averse firm will produce less under uncertainty than it would under certainty; the risk-neutral firm produces the same, and a risk-loving firm (possibly an empty box) would produce more. Investment decisions – choices of fixed capital, but also decisions about training a labor force – are affected by output price uncertainty, as well as by uncertainty regarding the future paths of input prices.

Some additional uncertainty situations involve the timing at which various decisions of the firm must be made. One timing issue deals with when output decisions must be made relative to when output prices are known, and the flexibility of firms in adjusting production in mid-stream. Another involves the timing of contracting for inputs – whether, say, capital must be chosen prior to knowledge of output prices but labor can be hired after those prices are known; not surprisingly, only if output can be chosen after labor is hired, is the labor flexibility valuable to the firm.

One of the issues of interest in this subject is the effect of greater risk rather than simply a comparison of certain and uncertain situations. A mean-preserving spread as the measure of increasing risk can affect optimal input ratios, depending on such technical characteristics of the production function as the elasticities of substitution between inputs and even on interactions among technical characteristics such as substitution elasticities and returns to scale.

7.6.2 *Search problems*

Search models have been developed to study problems in consumer and labor-market behavior. In each case, the problem is basically one of optimal information acquisition: when do you quit looking for a better deal? Models of consumer behavior typically involve search for what the consumer believes to be a “good” price. The search is costly, and more search costs more, so at some point in the search process, “enough is enough”: the consumer has sampled enough prices to collect a large enough sample to be confident that she’s not going to find a price that is better than the best one already found without paying more in additional search costs than the price difference is worth – if such a lower price exists at all. We could substitute quality for price,

and the problem would be basically the same. When the distribution of prices (qualities) is riskier, in the sense of a mean-preserving spread, the consumer actually stands a better chance of finding a lower price, given her unit search costs; in fact, in such a case, actual total search costs (unit cost times amount of search) will be lower than when the probability distribution of prices is tighter around the mean. One of the key concepts in search models is the reservation price or wage (in labor market search problems): the price the searchers have in mind such that, if they find a lower price, they stop the search immediately and buy. (Reverse the order of relative magnitude for a wage search.) The reservation price with which a searcher begins will be lower the lower her search costs are and the riskier the distribution of prices is (in the mean-preserving spread sense).

At any point in a search, the total search cost expended should not be of concern to the searcher: it is a sunk cost and therefore irrelevant to current decisions. However, the marginal expected return on the certain marginal search cost is quite relevant. What the previous search has accomplished is a greater fleshing out for the searcher of the outlines of the probability distribution of prices. The marginal probability that a lower price will be found as search continues will fall, eventually falling below either a constant or increasing marginal search cost.

7.7 **Suggestions for Using the Material of this Chapter**

Information was generally at a premium in antiquity, and risk (the absence of information) accordingly high. The organization of all sorts of transactions to protect participating parties from risk may have been pervasive, as has been well appreciated by historians of ancient agriculture such as Gallant (1991). I do not believe the extent to which such organization is visible in the archaeological record, even with a squint, has been examined. The contracts, implicit and explicit, involved in daily or even annual transactions will be largely invisible archaeologically – except possibly in preserved clay tablets found in the Near East and in papyri found in Egypt. The tablets from Kanesh may yield evidence of the implementation of principal–agent relationships

between business owners in Assyria and their kinsmen residing long-term in Anatolia.

Share renting contracts, which warrant more consideration than can be given them here, are one such case. Halstead relies implicitly on an incentive compatibility constraint in a putative sharecropping contract to offer an interpretation of the organization of production and harvesting of possibly palace-owned land in the Pylian Mycenaean kingdom (Halstead 1999, 322). Halstead's reasoning is that offering cultivators a share of the crop would have aligned their incentives with those of the palace, which was the effective owner, such as he admits ownership to be a relevant concept, of the land. He believes the sharing arrangement would have obviated the need for most if not all supervision of the cultivators, except, of course at threshing time, when he suggests the shares would be divided. Kehoe has given extensive attention to sharecropping (Kehoe 1988a, 558; 1988b 163–187; 1989, 40–42; 1997, 210–211); forward sales of crops and

even of offspring of animals and slaves (Kehoe 1989, 561–562; 1997, 211); partial remissions of receipts from purchasers of forward contracts in unfavorable years and other contractual devices for spreading risk both *ex ante* and *ex post* (2007, Chapter 3) in Roman agriculture during early Imperial times.

Much about information and risk involves learning (a flow concept), and production involves the application of knowledge (a stock concept), beliefs, and sometimes hopes, about how things work. The Bayesian approach to probability is a formalization of how learning changes information and beliefs, and the models of expectations formation offer guidance on the application of beliefs to actions, including consumption as well as production decisions. These models provide frameworks historians, philologists and archaeologists can use to analyze how experiences influenced actions in a variety of activities in antiquity, from farming to seafaring to science.

References

- Barnard, G.A., and Thomas Bayes. 1958. "Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances." *Biometrika* 45: 293–315.
- Bayes, Thomas. 1763. "An Essay toward Solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society of London* 53: 370–418.
- Conison, Alexander. 2012. *The Organization of Rome's Wine Trade*. Ph.D. dissertation. University of Michigan, Ann Arbor MI.
- Deming, W. Edwards. 1963. *Facsimiles of Two Papers by Bayes*. New York: Hafner.
- Descat, Raymond. 2011. "Labour in the Hellenistic Economy: Slavery as a Test Case." In *The Economies of Hellenistic Societies, Third to First Centuries BC*, edited by Zosia H. Archibald, John K. Davies, and Vincent Gabrielsen. Oxford: Oxford University Press, pp. 207–215.
- Fama, Eugene F. 1976. *Foundations of Finance; Portfolio Decisions and Securities Prices*. New York: Basic Books.
- Figueira, Thomas. 1986. "'Sitopolai' and 'Sitophylakes' in Lysias' 'Against the Graindealers': Governmental Intervention in the Athenian Economy." *Phoenix* 40: 149–171.
- Gallant, Thomas W. 1991. *Risk and Survival in Ancient Greece*. Stanford CA: Stanford University Press.
- Halstead, Paul. 1999. "Surplus and Sharecropper: The Grain Production Strategies of Mycenaean Palaces." In *Meletemata: Studies Presented to Malcolm Wiener as He Enters his 65th Year*, Aegaeum 20, edited by Philip P. Betancourt, Vassos Karageorghis, Robert Laffineur, and Wolf-Dieter Niemeier. Liège and Austin: Histoire de l'art et archeology de la Grèce antique, Université de Liège, and Program in Aegean Scripts and Prehistory, University of Texas at Austin, pp. 319–326.
- Hirshleifer, Jack, and John G. Riley. 1992. *The Analytics of Uncertainty and Information*. Cambridge: Cambridge University Press.
- Kehoe, Dennis P. 1988a. "Allocation of Risk and Investment on the Estates of Pliny the Younger." *Chiron* 18: 15–42.
- Kehoe, Dennis P. 1988b. *The Economics of Agriculture on Roman Imperial Estates in North Africa*. Hypomnemata 89. Göttingen: Vandenhoeck & Ruprecht.
- Kehoe, Dennis P. 1989. "Approaches to Economic Problems in the 'Letters' of Pliny the Younger. The Question of Risk in Agriculture." *Aufstieg und Niedergang der Römischen Welt II*. Berlin: De Gruyter, pp. 555–590.
- Kehoe, Dennis P. 1997. *Investment, Profit, and Tenancy; The Jurists and the Roman Agrarian Economy*. Ann Arbor MI: University of Michigan Press.

- Kehoe, Dennis P. 2007. *Law and the Rural Economy in the Roman Empire*. Ann Arbor MI: University of Michigan Press.
- Macho-Stadler, Inés, and J. David Pérez-Castrillo. 1997. *An Introduction to the Economics of Information; Incentives and Contracts*. Translated by Richard Watt. Oxford: Oxford University Press.
- McGuire, Robert A. 1980. "A Portfolio Analysis of Crop Diversification and Risk in the Cotton South." *Explorations in Economic History* 17: 342–371.
- Molho, Ian. 1997. *The Economics of Information; Lying and Cheating in Markets and Organizations*. Oxford: Blackwell.
- Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29: 315–335.
- Pratt, John W. 1995. "Foreword." In Louis Eeckhoudt and Christian Gollier, *Risk; Evaluation, Management and Sharing*, ix–xi. Translated by Val Lambson. New York: Harvester-Wheatsheaf.
- Salanié, Bernard. 1997. *The Economics of Contracts; A Primer*. Cambridge MA: MIT Press.
- Sharpe, William F. 1970. *Portfolio Theory and Capital Markets*. New York: McGraw-Hill.
- Sheffrin, Steven M. 1982. *Rational Expectations*. Cambridge: Cambridge University Press.
- Stroud, Ronald S. 1998. *The Athenian Grain-Tax Law of 374/3 B.C. Hesperia Supplements*, Vol. 29. Princeton: American School of Classical Studies at Athens.
- Westermann, William L. 1955. *The Slave Systems of Greek and Roman Antiquity*. Memoirs of the American Philosophical Society 40. Philadelphia: The American Philosophical Society.
- Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons, Inc.

Suggested Readings

- Hillier, Brian. 1997. *The Economics of Asymmetric Information*. New York: St. Martin's.
- Macho-Stadler, Inés, and David Pérez-Castrillo. 1997. *An Introduction to the Economics of Information; Incentives and Contracts*. Oxford: Oxford University Press.
- Pindyck, Robert S. and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapter 17.

Notes

- 1 But we can think of one: suppose you're a clothing manufacturer thinking of entering the market for children's clothes in another region or country. You have to make a decision about the array of sizes you produce. Are the nutritional conditions, as well as the genetic predispositions, of the population of the new market area the same as those in the market you've been serving? If not, you make an array of garment sizes that don't sell as well as you anticipated (that is, planned for when you were making your outlays).
- 2 Actually, a person who enjoys gambling might find it a challenge to discover the true odds of heads and tails with a dishonest coin, but the extent to which he or she is willing to assay the odds would depend on the degree of risk aversion and the sizes of the bets.
- 3 I'm playing a bit loose here for heuristic purposes. I have implicitly used continuous density functions, in which the probability of any particular outcome is zero, in the preceding text and in Figures 7.1 and 7.2. A discrete density function, which uses intervals (think of a bar graph that approximates the smooth curves of Figures 7.1 and 7.2), would yield a probability for events covered by a particular interval.
- 4 An interpretation originally made by Robert Schlaifer in a memo of November 13, 1961, to John Pratt, cited in Pratt (1995, x).
- 5 This exposition follows that of Hirshleifer and Riley (1992, 43–47).
- 6 Henceforth I adopt a notational convenience for the expression of the "difference of a difference" (which could be – and is – called a "second difference") $\Delta(\Delta v/\Delta c)/\Delta c$. In the place of this rather cumbersome notation, we will use v' to represent the first difference, $\Delta v/\Delta c$, and v'' to represent the second difference, $\Delta(\Delta v/\Delta c)/\Delta c$. Of course, the functions whose first and second differences we take will change – we may take first and second differences of, say the von Neumann–Morgenstern utility function over actions, $u(a)$, or first and second differences of production functions $Q = f(K, L)$, in which we will have to modify the notation somewhat to indicate which functional argument we are manipulating to get first and second differences of the function. Thus in the case of this production

- function, we will characterize the first difference of the function f (and hence of the output Q) caused by a small change in capital K as f_K and the second difference with respect to K as f_{KK} . Similarly, the first and second differences of f (or Q) with respect to labor L would be f_L and f_{LL} . Otherwise, our notation would continue to look like it did in Chapters 1 and 2: $\Delta(f/\Delta K)/\Delta K$, which, frankly, I find clumsy.
- 7 Named after Reverend Thomas Bayes' posthumously published paper, Bayes (1763); reprinted in Barnard and Bayes (1958) and in Deming (1963), which contains commentary by W. Edwards Deming.
 - 8 For purists, this is true only for independent probabilities. Temperature and humidity may not be independent, but for this example, we'll ignore any dependence. Thus, while hotness and wetness as represented by circles A and B are not mutually exclusive – they have an area of overlap – we do assume them to be independent in the sense that being hot does not affect the probability of being wet.
 - 9 The figure draws the curve for a distribution function as if the states were continuous. If there were a large number of states, this would be an approximation. With the four states in our example, a pure representation of the distribution of state probabilities would be a series of four bars or lines of different height. The continuous curve seemed to capture the intuition of the impact of new information on the prior beliefs better than did the discrete picture.
 - 10 Think about it this way for a minute: Have you ever asked someone you believe to be well informed on a particular subject their opinion regarding something about that subject, while holding some belief on the subject yourself? If their opinion, when offered, differed somewhat from yours, did you totally replace your own former opinion with theirs, or did you hold theirs alongside your own in case you needed to act on a belief (a posterior belief, now that you have some new information to put beside your own prior ideas) – so you could “think about it”? If you can recall such a situation, try to recall whether, when it came time to act on the matter, you used your own prior belief, your colleague's information which differed from your own belief (that is, you let them do your thinking for you; we hope that expressing it this way doesn't gum up the mental experiment), or did you use some combination of the two sets of information? We suspect that the latter of these three informational options is most likely to have been the case, particularly in actions that were serious enough to make information collecting and processing worthwhile.
 - 11 I say *may* because there is a further algebraic requirement. Each asset price can be broken down into a z_{as} -weighted sum of state-claim prices; this sum can be expressed in matrix notation as a vector of these yields (the z_{as}). The number of assets composed of *independent* yield vectors must be greater than or equal to the number of states. An independent vector is one that can't be obtained as a linear combination of any group of all the other yield vectors. For instance, suppose we have three yield vectors: $\mathbf{z}_1 = [z_{11} z_{21} z_{31}]$, $\mathbf{z}_2 = [z_{21} z_{22} z_{23}]$, and $\mathbf{z}_3 = [z_{31} z_{32} z_{33}]$. Can we get, say, vector $\mathbf{z}_1$ by adding together or subtracting the elements of the vectors $\mathbf{z}_2$ and $\mathbf{z}_3$, or some fractions or multiples of them? That is, is there some a and b such that $a\mathbf{z}_2 + b\mathbf{z}_3 = \mathbf{z}_1$, where that sum would look like $[(az_{21} + bz_{31})(az_{22} + bz_{32})(az_{23} + bz_{33})] = \mathbf{z}_1 = [z_{11} z_{21} z_{31}]$? If there isn't such a combination, then the vector of yields $\mathbf{z}_1$ is said to be independent of yield vectors $\mathbf{z}_2$ and $\mathbf{z}_3$. In this case in the asset problem, we would say that asset 1 could not be duplicated by buying a combination of assets 2 and 3, and therefore that the three assets are independent.
 - 12 The exposition below follows treatments in the following books: Sharpe (1970, particularly Appendix D); Fama (1976, especially Chapters 7 and 8) – both excellent texts, full of intuitive explanation; together with John Lintner, Sharpe won the Nobel Prize for the capital asset pricing model, described subsequently; and Hirshleifer and Riley (1992, 153–161), a more compact and consequently relatively less accessible treatment, but one which links the portfolio model with models of other behavior under uncertainty.
 - 13 If borrowing and lending are possible, the efficient set will extend beyond point h , representing “short selling,” either issuing a risky asset oneself or borrowing at an uncertain rate of interest.
 - 14 Two examples, one a contemporary scholar's inference, the other a likely consequence of a law in antiquity. First, as Kehoe (1988a, 33) suggested as a rule of thumb that aristocratic Roman landowners would have used with their tenants, “Thus Roman landowners dependent on tenants for their income had to find some way to induce their tenants to make investments that often would pay off only after the expiration of the original lease. Roman landlords might accomplish this purpose by taking advantage of the economic goals of their tenants.” Second, the case from antiquity: according to Stroud's (1998, 115) assessment of the Athenian grain-tax law of 374/3 B.C. (an 8 1/3%

tax in kind on grain from the islands of Skyros, Imbros, and Lemnos which was to be transported to the Aiakeion near the Athenian Agora and sold at auction, with the proceeds going to the military fund) the requirement that the (Athenian) tax farmers collecting this tax be responsible for transporting the grain from the islands to Athens had the incidental effect of encouraging them to bring additional grain on the ships, which would put into the Peiraeus to which the law specified the in-kind grain tax would be brought. This was during a period when all Athenian citizens shipping grain were required to bring those shipments to Athens (whose port was the Peiraeus) (Figueira 1986, 150). While this law did not affect

all Athenian shippers, a non-negligible share of Athens' grain imports came from these three islands, and the tax farmers, who probably were *emporoi*, would have profited from the tax farming and had the incentive to put additional grain on their ships carrying the in-kind grain tax.

- 15 Some writers appeal to four elements, separating what we refer to here as the third component into outcomes and payoffs.
- 16 The rational expectations model began with Muth (1961).
- 17 For an example of agricultural price determination under rational expectations, with inventories, see Sheffrin (1982, 165–166).

Capital

Most of the artifacts uncovered by archaeologists, reported by them, and long discussed thereafter by scholars of several disciplines are pieces of capital: buildings, equipment, and stores of value. When the subject is “the economy,” it is common to find detailed reports of implements, with analyses of what their uses might have been, their techniques of manufacture, and the means by which people figured out how to make them. In the literature on ships and shipping, the social relations surrounding those seagoing vehicles commonly are cast as, “Who owned the ship?” and “Who owned the goods carried?” Less commonly asked are, “How were the ships paid for?” and “Why would anyone want to own one, if they were so risky?”

The economics of capital studies the social relations between people and their long-lived, productive goods. These relationships continue over time, which introduces complications that have rendered capital theory one of the most difficult topics in economics. We offer just the barest introduction to the concepts used to study the resource allocation issues associated with capital. We open in sections 8.1 and 8.2 with a tour of basic concepts and continue in section 8.3 with an exposition of the concept of an interest rate. Section 8.4 offers a brief overview of the neoclassical theory of capital, an analytical

structure intended to account for the quantity of capital a society holds, how much it desires to hold, how rapidly it approaches the desired stock, and the magnitude of the rate of interest. Section 8.5 treats the investment decisions contained in section 8.4 with the greater realism made possible by focusing on the decisions of an individual producer. Section 8.6 offers the equivalent, zoom-lens view of the saving-consumption decision which permits the investment discussed in section 8.5.

8.1 The Substance and Concepts of Capital

The term “capital” has acquired so many meanings, in both colloquial usage and the technical lexicons of various disciplines of study that it will be fruitful to begin the discussion of it by enumerating as well as describing a bit what we will mean by the term in this exposition. There are also many analytical terms and concepts that are useful, even indispensable, for discussing capital from the perspective of resource allocation. After a relatively descriptive and taxonomic first subsection, we follow with six other subsections devoted to conceptual topics that arise in discussions of just about any aspect of capital.

8.1.1 *Capital as stuff*

In its abstract form, capital is a source of productivity, productive power, income. From a slightly different viewpoint, capital represents claims to income or consumption, in both the present and the future. Within this definitional scope, there are three principal categories of capital: material capital – buildings, tools, equipment, and so forth – human beings, and the stock of money. The major distinction between the first two categories is institutional, although in societies that permit slavery, much of that institutional distinction disappears. In those cases, there are sufficient physical differences between material and human capital to make the distinction worth retaining for analytical purposes, just as it frequently is useful to distinguish between buildings and equipment within the category of material capital. First is the mobility of labor – or human capital – relative to physical capital: it can be shifted among locations with considerable ease, at least when compared to buildings and large, heavy, sometimes fragile pieces of equipment. Second is the flexibility of its use. The application of labor to a particular productive activity can be varied literally by the minute or hour, whereas with many forms of material capital you're pretty well stuck with what you install for a longer time. For instance, if you make a bronze spearhead and then decide you wish you had made four arrowheads or a plowshare instead, you *can* smelt down the spearhead and recast it, but it probably would be cheaper, as well as faster (part of "cheapness"), to find somebody who's willing to part with the other capital items on some acceptable terms.

Third, there is the matter of what might be called the deferability of capital costs and maintenance. If you don't maintain a particular machine as well as you know you should – say, because you're strapped for cash, so to speak – metal gears might rust or corrode, a roof might develop an untimely leak, and so on, but in many cases the equipment will continue to function, and if it does fail – say, a shovel handle or blade breaks, or ropes fray and break – you can repair it in many cases.¹ Labor – and we're thinking particularly of slave or some closely related type of dependent labor – on the other hand, if you don't feed it

regularly or shelter it, can die, irreversibly. The replaceability of such "failed" human capital depends on its supply conditions, and from the owner's perspective, it may be so plentiful that it's cheaper to let it die than care for it as would befit a good shovel. The Laureion mine slaves of the fifth and fourth centuries B.C.E., and possibly Roman agricultural slaves of the early Imperial Period, are examples of such low-maintenance human capital. There are many intricate and important issues in slavery involving supply conditions, but we will defer our principal treatment of them until the chapter on labor.

Fourth, human capital appears long to have demonstrated a mind of its own, despite some parts of the literature that emphasize docility and even tendencies toward the autonomic on the part of free labor. The Romans faced some famous and extensive slave revolts and the Greeks some lesser ones (Westermann 1955, 41–42, 64–65, 75), but short of outright revolt, slaves may be in positions to negotiate some aspects of their work and nonwork lives with their owners in ways that do not exist for inanimate equipment (McKeown 2011).² The substitution of material capital for slave capital may sometimes be profitable simply because the slaves are too much trouble. While it may pay the owner of a metal tool to lock it up at night, the costs of guarding slave capital may be more extensive, "protecting" it against forces "inside the corral" as well as forces outside.

Fifth, and final in our list of reasons to distinguish between material and human capital, is the moral distinction, which affects both the persons involved in the actual slave system in antiquity and the modern students of those activities. We list this issue last simply because much – but not necessarily all – of the thought that would inform its consideration lies outside the contributions of economic theory. Modern students of antiquity find themselves faced with opportunities for moral judgments of ancient individuals and groups that current ethical traditions find it difficult to avoid. Sometimes such considerations may affect the analytical apparatuses chosen for investigation, and even the topics. This interaction between the ethical norms of students and their subjects, both actual and potential, must be kept in mind rather than dodged. On the other

hand, the ancients themselves occasionally wrote down their thoughts regarding ownership and use of their fellow humans, and even when their conclusions justified themselves in their own eyes, the very fact that they thought about it at all indicates that it was indeed an issue distinct in their own minds from the use of shovels (Westermann 1955, 40, 76, 79., 81, 106, 113, 116).

Having rushed into a discussion of labor as capital, let's return a moment to some of the more obvious forms of capital – equipment and buildings and closely related items. A stone tool, to the extent that it is expected to yield productive service over some time period, is a piece of capital. If some stone can be picked up and used with literally no shaping – you didn't even have to spend much time looking for it – it is still a capital implement. There is some interest to the fact that its production appears to be costless; possibly more relevant from the cost perspective would be its replacement cost: would you expect to be able to pick up a perfect substitute for it at your feet just as easily as you found the first rock if you lose or crack the first one? Let's proceed to stone implements that definitely require some shaping: querns, scrapers, arrow and spear heads, mortars and pestles, stone vases; these are definitely costly, and they are expected to provide service over time. You might pause for a minute at the concept of arrow and spear heads (particularly arrow heads) providing service over time if what you use them for mostly is shooting at enemies from whom you don't expect to retrieve them. In such cases, you use them only once, so what's the deal about "continued service"? We'll grant, this could be a case for splitting hairs, but let's discuss some aspects of these implements. First, they are produced goods – you have to make them or acquire them from someone who did. Second, they are not final consumption goods – you can't eat them, wear them, or whatever. Third, this means that they are valued only for the service they provide – killing, protecting, feeding, offering prestige. Fourth, they frequently are carried in groups over some time period in which their user (carrier) may or may not have to use them, but when the need for one or more of them arises, they're really nice to have. Thus, they're carried in productive

inventories. Is a contemporary bullet for use in a rifle properly called capital or something else? The answer anyone gives to this question is probably less interesting than the reflections regarding the relationships between the bullet and its user.

Let's skip buildings and the really obvious tools like agricultural and construction implements for the moment and pursue the concept of inventories as a category of capital in addition to equipment and buildings. Having introduced inventories with an example from military stores, let's think of inventories of consumer goods. Inventories are stocks of something that are held for future use, either in consumption or in production, as contrasted to immediate sale or consumption. Holding them provides a service, and eventually discharging items from an inventory and consuming them – in either production or consumption – provides a future service. Inventories are a category of capital. When an inventory happens to be a consumption good, such as food or cloth, both capital and consumption goods happen to be of the same "stuff." This may seem to be the first order of obvious, but when we are searching for simplifications of the capital concept below so we can develop models of capital that are simple enough to understand (models of capital get remarkably complicated very quickly, requiring radical simplifications to yield insights), one of the best simplifications available is to think of the capital good and the consumption good as the same thing, just used for different purposes (immediate consumption or deferred consumption – that is, production). In general, inventories can contain either production goods (inputs such as seed, spare tools, various materials) or consumption goods (held by either producers for sale at opportune times or by consumers for consumption or resale at convenient times). In both cases, these inventories are part of the society's capital stock. Their values may change more sharply – and predictably – over various time periods than do those of other capital in the form of equipment and buildings, but that does not detract from their status as capital. The periodicity of use of some goods does not correspond well to the opportunities for acquiring them, and the particular service conferred

by inventories is to save on predictable transactions costs (and sometimes unpredictable ones as well). Some production processes routinely apply labor and some materials to inventories of raw or semi-finished materials in a continuous fashion. We could think of household stores of food grains, oils, dried fruits, and so forth similarly, as inputs routinely used in a major household production activity – production and consumption of meals.

Let's go back and pick up buildings and equipment now. Buildings, both private and public, probably comprised the largest single component of the societal capital stocks in the ancient Near East and Aegean. It may be tempting to believe that public buildings claimed a much larger share of the building capital stock in those societies, particularly in Egypt and the Mesopotamian cities, primarily because of the durability of many of those monumental structures. We frequently know much less about domestic architecture than about the monumental buildings – somewhat less regarding the construction and cost of what might be a typical domestic building, but very much less about the number of such buildings in antiquity. In bringing up the subject of the share of total building stocks accounted for by public and private structures, we must confront the methodology of such comparisons: how to add the Ramesseum to the inferred private houses from the remains of, say, four domestic buildings in Deir el-Medina? This may seem impossible to many and downright offensive to some, without throwing in the capital stock in productive and military equipment which, of course, would have to be included in a measure – either empirical or remaining at the conceptual level – of the entire society's capital stock. Empirically, measurement of a capital stock in a contemporary economy is difficult, and we have no intention of suggesting actual measurement of the capital stock of an ancient economy – most of which has long disappeared – but it is a useful thing to think about. We will deal with measurement of capital stocks in section 8.1.6, but we turn to several other capital forms before departing this section.

Before departing the public capital stock with discussion of buildings alone, we must point also to roads, harbor and river improvements, canals,

bridges, city walls, water supply systems including wells and aqueducts, and undoubtedly other structures. These are all part of the public (as contrasted with the private) capital stock. Absent extensive private investment in production facilities (factories and warehouses), and considering the possibility of relatively flimsy construction (that is, small quantities of capital) of domestic structures, the public component of the capital stocks of many of these ancient societies might have been relatively large.

Funeral architecture is largely excluded from contemporary accounts of capital stocks but it might be useful to include them in ancient accounts, although the concept of a service in the case of those structures may require some rearrangement. Alternatively, a case might be made that some portion at least of the durable structures and goods devoted to funerary purposes be stricken from the “books” of useful capital stock. Either case could be argued, insights could be gained from both sides of the argument, and there might be no reason to actually decide in favor of one or the other. Certainly the buildings devoted to tombs and related structures did not produce consumption goods for the living citizens, and the durable goods buried in them ceased to produce service flows. On the other hand, in some of these societies, funerary activities appear to have created considerable employment, and to say that these structures produced no flow of services to the living may be to misunderstand the relationship between the spiritual and physical lives of the people.³

Vehicles, ships, weaponry, tools for manufacturing, construction and agriculture – these are all well reported in specialized articles and monographs that describe how they appear to have been made and sometimes the metallurgical and mechanical technologies involved in their construction and use (for example, vehicles: Crouwel 1981, 1992; ships: Steffy 1994; weaponry: Buchholz and Wiesner 1977, Aldrete *et al.* 2013; manufacturing and construction tools: Olesen 2008, Parts III and IV; agricultural tools: Isager and Skydsgaard 1992, Chapter 3). They are capital. As we will see below, they contain interest rates implicitly, and they represent deferrals of current consumption in favor of future consumption. In

some instances, they embody the latest technological knowledge of their societies, particularly the weaponry. They represent investments that conferred returns on their owners. These are not the most common perspectives from which these artifacts are studied, but these characterizations emphasize their roles in the economies of the time rather than their physical characteristics, which have been traditional objects of study.

Now, what about capital in the form of a stock of money? We will consider money, and the banking industry that produces (supplies) it, in a separate chapter, but a society's stock of money is indeed a part of its capital stock inasmuch as it represents socially approved claims to consumption. To discuss the place of a money stock in an economy's capital stock we need to introduce a little additional terminology that we will explore further in the money and banking chapter, specifically a "commodity standard" for money. The major distinction is between a commodity money and a fiat money, the latter being composed of essentially worthless items, the former comprising a set of objects that actually have some alternative uses. The best example of a fiat money stock would, of course, be paper money, which was largely unknown in antiquity.⁴ Commodity standards include monetary systems that use, say, stones, for which the recent Yap Islanders are famous among anthropologists; metals, precious and otherwise; shells such as cowries; and consumption goods such as grains or animals. An important economic distinction between fiat and commodity moneys is the amount of resources tied up in the money stock. A fiat system uses an inconsequential proportion of its society's resources, but a commodity monetary system can require considerable resources that could be used otherwise for direct consumption or as productive resources capable of producing consumption goods. The resources used to produce or acquire the commodity money may be substantial: metals must be mined; stones or shells must be acquired through quarrying or trade or both; consumption goods such as grains, if stocks of them are indeed held instead of their being used simply as an accounting system without any stock holding beyond normal consumption inventories, must be produced and storage facilities for them built.⁵

Having expounded on commodity versus fiat monetary standards at some length, we should come to our purpose in distinguishing between money and the other categories of capital. The services provided by material and human capital stocks vary in some generally positive proportion to the size of the stock, but with a fiat monetary standard, the relationship between services provided and the size of the stock is more intricate, and in some senses the services are provided merely by the existence of the stock, with variations in the size of the stock being responsible for many other important consequences. To some extent, the same is true of the stock of a commodity money: more of it may simply redenominate the prices of all goods and services in the economy rather than provide directly for the creation of additional goods and services.⁶

8.1.2 *Capital in the production function*

While there are different categories of capital, there are also different ways to conceptualize capital as a type of stuff and alternative ways to think about how capital produces things. In some cases there really are different forms of capital, such that one way to conceptualize a particular capital item really is correct, or at least better suited than another way. Sometimes different perspectives simply can illuminate different facets of a capital item. In this subsection we describe three principal conceptualizations of capital and several models of how capital operates through time in productive processes.

The views of capital in capital theory fall into three major categories: an inventory concept of capital, in which physical capital is a stock of finished, consumable goods; a goods-in-process concept, in which there is a time structure of degrees of completion of various individual units; and the durable good concept, in which capital is a distinct, productive commodity. Each conceptualization (model) has had an important role in the development of capital theory, and each has been useful for helping understand particular aspects of capital. The inventory concept has been particularly useful in studying the basic relationships of capital theory – the choice between consumption and saving or between

consumption at different times, the accumulation of capital over time through investment, and the determination of interest rates – without entering into the difficulties of the age structure of the capital stock. We will demonstrate some of these concepts with a particular model of this genre in section 8.4.

The goods-in-process model is illustrated by a stand of trees of different ages. The growth rates of the trees as timber depend on their age, so the growth rate of the entire stand depends on the age composition of the trees. The concept of the “period of production” derives from this view of capital. The same quantity of capital – say, two stands of trees that are allowed to mature to different ages before cutting – can be associated with different equilibrium rates of output (cut timber), depending on its age structure.

The durable goods model of capital is the concept used most commonly today in analyses of the productivity of capital equipment and investment. A stock of capital represented by a durable good such as a machine, or a building yields a flow of productive services over a technical or economic lifetime. This concept of capital most closely corresponds to the heterogeneity of capital equipment in most economies, modern and ancient, and in this capacity is quite useful in the production function apparatus. Nonetheless, there are technical difficulties in aggregating measures of such heterogeneous capital equipment into larger scale “sectors” of an economy for analytical purposes. These aggregation problems have proven difficult to solve exactly, and most of the economics profession has agreed to live with the inexactness while some scholars search for proper solutions, on the grounds that (i) the worst possible consequences of the aggregation problems are not highly likely, empirically, (ii) all theory involves some inexactness, and (iii) there are no practical alternatives.

Within the scope of these three concepts of capital, the temporal structure of input-output relationships in production likewise has found a taxonomy that expands on the static structure of capital and capital services that we used in Chapter 1. The first two structures are associated with the goods-in-process vision of capital. In the “point input-point output” production structure,

all inputs are applied at a single moment, and all outputs appear at a single moment some variable time later. In the “continuous input-point output” model, inputs are applied continuously while outputs emerge at discrete intervals. The ageing of grape juice into wine is a prototypical example of this time structure. Typical of the durable goods concept of capital is the “point input-continuous output” structure. An input is put in place at a specific date, and it continues to yield outputs in a continuous or periodic stream over time. The inventory view of capital typically has a “continuous input-continuous output” time structure of its input-output relations.

8.1.3 *Stocks, flows, and accumulation*

We referred in Chapter 1 to the distinction between stocks and flows, but the present chapter is an appropriate place to expand on it. A flow, be it of capital services, labor services, water, consumption, or whatever, is a quantity during a specific, delimited time. It has the dimensions of a quantity employed per unit of time. A stock is a quantity in existence at a specific point in time. We can think of a stock existing at multiple points in time, but each stock refers to exactly one of those times. All flows come from some stock, again whether the stock and corresponding flow refer to capital, labor or some consumption goods.

The size of a stock may be altered, either deliberately or inadvertently. Such changes are accomplished by means of flows of the same substance as the stock. A stock of capital can be increased by a flow of investment or decreased by a flow of deterioration in excess of what investment is able to replace. The concept of equilibrium we have used till now – we have not made a big deal of the concept, but it has been a constant, analytical workhorse – has been essentially timeless, because most of our analysis has abstracted from the passage of time. This avoidance of such an obvious fact of life may seem silly, even absurd to some, but it is a device for focusing on an eventual conjunction of quantities and prices that will satisfy all the relationships involved in a particular resource allocation problem. This timeless equilibrium concept is called “static equilibrium.”

It is contrasted with equilibrium concepts that take explicit allowance of the passage of time, with the behavioral links between neighboring time periods explicitly delineated – and called, naturally, “dynamic equilibrium.” Once we introduce stocks and time to the static flow analysis, we encounter the possibility that, while all the flows may be in equilibrium during each and every time period, as long as investment flows are nonzero (either positive or negative), it will be very difficult to devise an equilibrium concept that will make the stock of capital be “in equilibrium” in each time period. Only when net investment (that is, anything beyond replacement investment) is zero will we have the equilibrium stock of capital.

Reflecting this thin end of the wedge of the enormous complications that time introduces to capital theory and to dynamic (that is, intertemporal) economic analysis in general, there are several analytical baselines used in capital theory that have not appeared in our static references so far. First is what is called the “stationary state,” a situation enduring over time in which all production and consumption variables remain at the same value. There is no growth. Next is “steady-state growth” (sometimes called “semi-stationary growth”), in which all variables grow at a constant proportional rate, and the economy is the same over time except for a change in scale. An increasing or decreasing growth rate would not satisfy steady-state growth. These benchmarks are *not* actual, observed phenomena, but are simplifying analytical references used to help isolate problems for study. We will encounter these terms more in the chapter on growth than in the present chapter, but we will refer to steady-state growth momentarily in section 8.4 to discuss relationships between the abundance of capital, the price of capital, and the interest rate.

Before departing stocks, flows and accumulation, we emphasize the way we will use the term “investment” here. We are not thinking of just trading around a fixed quantity of capital, such that some producers obtain more and others use less. It is common in colloquial terminology to think of, say, buying the stocks and bonds of corporations as investing. In fact, that is saving, whereas investment is what the company does with the funds you let it use when you buy the

stocks. Investment as we will think of it here involves changing the total quantity of capital, not just the composition of one user’s quantity.

8.1.4 *Prices and values*

In the static treatment of capital services we introduced the pricing and valuation concepts appropriate to capital: rental rates, interest rates, marginal products, and capital goods prices. Rather than simply review these concepts, we will extend them a bit. The value of marginal product (or “marginal value product”) of capital is the marginal physical product of capital services times the price of the product the capital equipment is used to produce. Just as the colloquial term “wage rate” has been attached to the marginal value product of labor whether the suppliers of labor think of themselves as wage workers or not, the marginal value product of capital has been pretty much irretrievably dubbed “the rental rate” of capital. You may think of the rental rate on an apartment you once occupied and wonder how that payment compared with the rental rate used as a technical term. It’s a pretty close correspondence, in fact. You were paying for the services of the part of the building represented by your apartment. In a town of any size there should have been sufficient competition among landlords to drive most of the pure profit out of most tenants’ rental payments (regardless of popular suspicions of landlords in university areas). Some taxes might have been included in the monthly rent bill, and you might have had some utilities included, but other than that, what you paid was pretty close to the theoretical concept of the rental rate on the quantity of apartment you occupied. When we introduce the concept of the quasi-rent in section 8.2, we will introduce the possibility that the landlord picked up a slight amount more than the marginal product of a perfectly variable factor by virtue of the fact that once the building was constructed, its quantity could be varied only over time through net investment or through insufficient replacement investment to maintain it in its original condition.

As what you paid in your monthly rent was, let us concede roughly, one-twelfth of the annual marginal product, how much should a reasonably bright person have been willing to pay for the

entire building (say, in a transfer of property between slumlords)? Let's identify the monthly rent on all the apartments in the building as r and the annual rent bill as R and just work from now on in terms of annual rents on this building. Thus the building pays its owner the amount R this year. Let's ignore depreciation (and hence maintenance), taxes, and all the uncertainties that can go along with urban building ownership; next year the owner can expect to earn R off the building too, but the value today of next year's income must be discounted (we'll get to more on why in section 8.3 on interest rates). And similarly over all the ensuing years. The present value of the income stream is $R_0 + R_1/(1+i) + R_2/(1+i)^2 + \dots + R_n/(1+i)^n$, where i is the interest rate, and the subscripts on the annual rentals refer to years.⁷ The annuity formula for this sum is $(R/i)[1 - 1/(1+i)^n]$; as long as the interest rate is positive, and if the number of years over which the building can be expected to render services is very long, this formula simplifies approximately to R/i . This being the present value of the income the building will deliver over its expected lifetime (in terms of either the cash or beans you offer your landlord monthly or the value of the shelter services its occupants receive over that period), it is the price that the reasonably bright buyer would be willing to offer for it. Designating the price of this capital asset as P , we have the relationship $P = R/i$ (approximately).⁸

A slight variant on this concept of the price of an income stream will be used in the exposition of capital theory in section 8.4, but this is an excellent place to introduce it. We have used the concept of an income stream in deriving the price of a capital asset in the previous paragraph; it is just the amount R , in each of the future time periods over which the asset is expected to continue in production. We can conceive of an asset that we could call a "permanent income stream," which delivered some amount of income R^* indefinitely. We could think of buying or selling permanent income streams of different sizes (that is, $R^* = 1$ deben per year, $R^* = 2$ deben per year, $R^* = 5$ deben per year, and so on, of whatever size, and in whatever numeraire – metals, bags of lentils – we wanted). The price of a one-unit permanent income stream is the inverse of the interest rate: $1/i$ ($R^* = 1$, so $P = 1/i$).

Now, let's think about the ratios of factor payments like the wage rate and the rental rate. From our discussion in Chapters 1, 2, and 5, we developed the expectation that the ratio of factor quantities used in production will vary in inverse proportion to the ratios of their factor prices (where we use the term "factor price" to refer to the prices of the service flows rather than of the stocks). The rental rate on capital and the wage rate on labor have similar dimensions: if we use a monetary measure of product value, it is shekels per labor service unit (hour, day) and shekels per capital service flow unit (say, an hour's or day's service of a certain quantity of capital). Divide the wage rate by the rental rate, the shekels cancel and we get hours of capital service divided by hours of labor – physical units entirely. If all prices were to rise or fall by the same proportion, the wage / rental ratio would be unaffected since the price terms have canceled. The comparable ratio of the wage rate and the interest rate does not lend itself to such a neat result because the interest rate itself is a pure number (that is, without any physical dimensions), and the value of that ratio *would* be affected by a proportional change in all prices as will be discussed in Chapter 9.

8.1.5 Temporal aspects of capital

In this subsection we introduce some concepts that involve relationships between capital at different times. The three topics we address here are durability of capital equipment, the ageing of capital, and the embodiment in capital of technology and technological change.

Durability

In our emphasis to date on the flow of services over time from a piece of capital, the issue of how long a specific unit of capital can continue to deliver those flows has taken a back seat. The durability of a piece of capital is a choice rather than a technologically determined datum. The choice about durability is separate from the productivity of the same piece of capital: durability does not affect productivity at any point in time but instead characterizes how long particular levels of productivity can be expected to last. The typical measure of durability is the expected or planned life of a piece of capital, and bringing up the subject of the life of a piece of equipment

raises the issues of equipment replacement (when to replace?) and scrapping (you scrap an old piece of capital just before you replace it with a new unit).

In planning the durability of a piece of equipment or a building prior to constructing it (or contracting for its construction or looking around at what's available to pick the "best" durability for his purposes), the tradeoff the user makes is between an increasing cost of greater durability and what that increased durability buys him. To complicate matters somewhat, what durability of capital buys a user of capital depends on the kind of planning horizon the user has. If he is thinking in terms of a single, say, machine or building, without replacing it at the end of its life, there is one sort of problem to be solved by optimizing durability. If, on the other hand, this user of capital – we could call it a firm for simplicity – has in mind replacing the equipment at the end of its life, and replacing the replacement machine at the end of its life, and so on *ad infinitum*, changing durability confers another type of benefit. In the former case, when the lifetime of a single, nonreplaced machine or building is being considered, the marginal cost of providing another period of durability when constructing the capital is equated to the present discounted value of the return the machine will deliver in that extra period of useful life. In this case, a higher interest rate (at which present values are discounted) will reduce durability, a higher product price will increase durability, and higher input prices will shorten optimum durability.

In the other case, in which the current capital investment is just the first in a very long, anticipated chain of replacements, the benefit of extending durability is the ability to defer replacing the capital at the end of its useful economic life. Thus, in terms of equating the marginal cost of additional durability with the marginal benefit thereof, the optimal durability of a machine will be found where the cost of lengthening durability for, say, one period equals the interest for that period on the present discounted value of costs of all future replacements. Lengthening durability pays as long as the cost is less than this interest saving. Depending on the planning period of the

firm installing and using the capital, the value of the output may not affect optimal durability, and the problem boils down to one of simple cost minimization by choosing optimal durability. Other input costs may affect both sides of the durability $MC = MB$ relationship (MB is marginal benefit, or the extra benefit from increased durability), but it is difficult to predict in general how those costs will affect the durability (measured in length of equipment life) that equalizes MC and MB. A higher interest rate will, in this case as well, shorten optimal durability.

Ageing of capital

Most of the topics in the ageing of capital equipment tend to be on the sad side, as there are few compensatory benefits of ageing capital as there are with ageing labor. We deal with deterioration and decay, obsolescence, and depreciation. Actually there is quite a bit to say about each of these topics. Taking deterioration and decay first, there are alternative models of those processes which force us to think about how we believe a particular type of capital will change over time. A very simple model of capital deterioration is known as the "one hoss shay" model: the capital equipment experiences constant efficiency over its lifetime but at the end of its lifetime it falls completely apart. The certainty version of the one hoss shay model is handy in the study of certain types of problems. Somewhat more realistic is the stochastic version of the one hoss shay, in which the capital equipment has constant efficiency over its lifetime but also has a probability of failure, which may be either constant or exponential (that is, the failure probability increases over time). Among contemporary technologies, light bulbs are examples of the one hoss shay deterioration model, with either constant or exponential failure probability.

The other principal model of capital deterioration doesn't have as catchy a name: you can just call it the declining efficiency model. Declining efficiency can take a number of forms that have different economic implications. It can be nothing more complicated than a smaller number of units of output per unit time of capital operation, over time. Alternatively, declining efficiency could take the form of a machine requiring more

coordinating inputs (say, labor or materials) to produce a constant rate of output as the equipment ages. In this case, while the gross output may remain constant, the net output declines. Producers using capital of this sort are indifferent between units of different ages ("vintages") because they produce the same output, just in different quantities. More precisely, a capital-using firm would be indifferent between having one unit of capital built, say, two periods ago and $(1-\delta)^2$ units of brand new capital, where δ is the annual decay rate, and the squaring of the $1-\delta$ indicates that the decay occurs for two periods. In this case, the outputs of the capital of different vintages are perfect substitutes. A different declining efficiency model would have capital of different ages produce output of differing quality; in this case, as consumers are not indifferent between the outputs of the different ages of equipment, producers likewise are not indifferent between deterioration-adjusted quantities of the capital.

As with durability, deterioration is an economic choice. In fact, the rate of deterioration reflects, in part, the prior economic choice of durability: all other things being the same (how the capital is used and treated), more durable equipment will have a lower deterioration rate than less durable capital. Two other economic decisions also affect deterioration, however: maintenance and intensity of use. Intensity of use (sometimes called capacity utilization) is rather difficult to define, but it is possible to specify conditions that would represent more intensive use. The "normal" operating hours of a piece of equipment would include some down time for maintenance, and the number of shifts of labor operating it. If the demand for the services of some equipment were temporarily higher than normal, some routine maintenance time might be deferred, and the equipment might be operated for a longer time per day, week, or month for a while. Such operation could be expected to accelerate the deterioration of the equipment – that is, use it up more rapidly. Maintenance also affects the rate of deterioration of capital, and the tradeoff the capital user faces in deciding how much maintenance to perform is the value of the deferred deterioration obtained per unit of maintenance cost.

At the margin, these two will be equalized in an optimal maintenance program. Again, we should remind the reader that in societies without a lot of sophisticated accounting and record keeping, we recognize that "optimal" anything – maintenance schedules, durability choices, whatever – are going to look quite a bit different "on the ground" than they do on the mathematical economist's note pad or blackboard, but we equally believe that these concepts would have guided the users of capital as much four or five thousand years ago as they guide large corporations today. An ancient Egyptian could be expected to look at a shovel or a plough sometime early in its life and ask himself, "Do I straighten it out, oil it down (or whatever the appropriate actions might have been), or do I just let it go?" Late in the equipment's life, he could be expected to have asked himself, "Do I repair it or just scrap it and get a new one?" The factors entering his calculations, implicit as many of them may have been, are the ones we have described.

Another thing happens to capital over time, even if absolutely no efficiency deterioration or increased probability of failure occurs. If newer, more efficient equipment is introduced, the older equipment becomes obsolete. The value of the old equipment will fall in proportion to the relative efficiency improvement of the new capital.

Depreciation is the price reflection of the combination of deterioration and obsolescence. The decline in the market value of capital equipment over time is called "depreciation." Some readers may object that second-hand capital markets certainly did not exist in the ancient Mediterranean societies. They are highly imperfect in today's industrialized societies too, but that does not affect the fact of depreciation, only the ability to measure it. Users of capital will behave according to their measures of depreciation, be they rough and ready or sophisticated.

Any capital stock has an age composition, which is simply a dressed-up way of saying that different pieces of equipment (or buildings) in some collection of capital goods were built at different times. (Keep in mind that capital goods of different ages could be acquired at the same time, but that would not make those goods "the same age" in any relevant, economic sense.) For a

little specificity, let's suppose that we're referring to the capital stock of a large, independent producer, such as one of the major, Egyptian temple complexes at Thebes during the New Kingdom, or a large, private, Roman oil producing family firm of the second to first centuries B.C.E. The interest rate only indirectly affects the age composition of a capital stock. At any point in time, one of these producers has a variety of capital goods of different ages. If the interest rate were to rise, all capital good prices would fall but the older units of capital would increase in value relative to the newer goods because they don't have as much future yield to be discounted at the higher interest rate as do the newer items. This relative valuation effect is larger the greater is the durability of capital and the higher is the interest rate. The more durable is the equipment on average, given the age of any particular piece of equipment, the larger will be the price reaction. If a producer expected the price of new capital to remain constant while the interest rate rose (possibly because the production unit imports its capital from an area unaffected by the interest rate increase), that firm would find it to its advantage to hold old capital. Alternatively, if the producing firm expected the price of capital goods to rise, its user cost of capital (discussed in more detail below in section 8.5) would fall because of the capital gain. In this case the user cost of newer equipment would be decreased by more than that of older equipment, this time because the longer future yield stream of the newer equipment offers greater scope for capital gain.

"Embodiment"

How does technological progress get transmitted from minds and drawing boards to increases in productive capacity? One way of looking at this problem is to think of new productive knowledge as being either embodied in new capital or being "disembodied," which has less precise characterizations than does the embodiment hypothesis. Disembodied technological change would include new ways of organizing and using existing equipment and labor. With embodied technological progress, the new knowledge is ineffective until it is embodied in new capital equipment that employs the improved principles. With this type of technical change, turnover of

the capital stock through investment is the means of raising productivity. Clearly there is room for both methods of entry of new knowledge into the economy. If embodied technical change reaches a certain pace, investors will not get enough time to repay their investments in new technology before even newer technology becomes available and depresses the return on the previous investments in new technology. As a result, investment will be retarded by the pace of technological change. Such rapidity of technical advance is unlikely to have characterized many times and places in the ancient Mediterranean region, but the notion is worth keeping in mind. This distinction between technical progress mechanisms may offer some perspectives from which to investigate technological change in the ancient Mediterranean societies – the role of investment in changing technology as well as just increasing the quantity of capital, the importance of capital equipment in transmitting knowledge, the remaining evidence of organizational (or "soft") technologies in enhancing productivity and material wellbeing.

8.1.6 Measuring capital

One reason we, as analysts, might want to measure capital is to use it as a factor of production in a production function analysis. Other models of economic decisions rely on the quantity or value of capital as an explanatory variable. Yet our example of finding a way to "measure" the Ramesseum and some private dwellings in some mutually compatible fashion certainly alerted the reader to the degree of heterogeneity that measurement of a capital stock must accommodate. The first, and most obvious element of heterogeneity involved in measuring capital is that capital takes a wide variety of forms. The Ramesseum and the private dwellings are at least all buildings. Some measures of capital would include stone, wooden and metal tools, public roadways, and inventories of grain as well as the Ramesseum and many other buildings, humble and magnificent. The second dimension of heterogeneity involves the age of units of capital that are otherwise quite similar. As we noted in section 8.1.5, capital of different ages differs in productivity at the time of measurement and, regardless of relative productivities, may be expected to continue to offer services

for different lengths of time in the future. For some purposes, a measurement of capital stock that accounted for different remaining lifetimes would be useful, but for an analysis of current productivity, a capital unit's expected services in the future are irrelevant. A durable shovel may be just as productive this year as one that will fall apart at the end of this year. However, if we wanted to capture something about the wealth of the society or some members of it, it would be quite useful to distinguish between shovels that will last several more years and ones that will end their lives this year. For purposes of use in a production function, we want an index that reflects the physical quantity of capital in use, even if we are forced to measure it in value terms because of its heterogeneity. Fortunately, there are alternative, equally correct methods of measuring capital stocks, either for all of society or for more restricted sectors of it.

The two principal perspectives that can be taken in measuring capital are book value and market value. Both use capital good prices as the means of bridging the physical heterogeneity of the vastly different forms in which capital is found. Book value is a retrospective measure, while market value is forward looking. In practice, the methods of getting the original data for either construct involve considerable intricacies, most of which we need not enter here. Constructing a book-value measure of capital stock would involve finding the original price (cost) of a currently existing capital good, converting that original price to current prices with an appropriate price index, allowing for depreciation (both physical deterioration and obsolescence), and deriving a current-period-equivalent valuation of the remaining productive capital embodied in the unit. This may involve going back nearly a century, depending on the type of capital being measured. A variant on this method is to measure replacement cost, one complication of which is that current production technology may offer a different cost structure of building the piece of capital under question than the technology that originally built it. The perpetual inventory method is yet another method of implementing the book-value measure of a capital stock: start with an index of capital value at some date in the past, adjust for price-level differences with the appropriate price index, and add each year's

investment less depreciation from the base year to the present.

The market-value approach to measuring capital stock is forward looking because the current valuation of a unit of capital is based on the future income stream it is expected to yield. Older equipment with a shorter remaining economic lifetime will be discounted accordingly relative to comparable but newer equipment. Less productive equipment of identical age will be valued proportionately to more productive equipment. When capital takes different forms – buildings, vehicles, tools – each will be valued according to the incomes it is expected to yield over its economic life. Issues of replacement cost and the proper depreciation formula to use are moot in the market-value approach. The market-value approach will be sensitive to the prevailing array of interest rates at the time of measurement, a higher interest rate structure tending to depress capital good prices.

We must remember that a capital stock is measured at a single point in time. In even thinking about ancient capital stocks, and particularly when attempting to measure them, we should pay attention to what items we think were in use at the time we have in mind and which had been discarded and were effectively out of the capital stock for the date we are contemplating. As to the extent of aggregation, for some purposes, it might be perfectly satisfactory simply to add the number of, say, plough shares or hoes available to particular individuals or groups at a particular place and time. We need not always aggregate all the different forms of capital into a single, aggregate measure of capital. For instance, when studying agricultural production in some region and time, we might decide to think in terms of several distinct capital goods: buildings, tools, vehicles, animals. While we would have four aggregation problems instead of one, each of the four might be far more manageable in its own right than would a total aggregation across the categories. And besides, we might want to know about the relative productive contributions of the capital stocks in these four categories.

8.1.7 *The labor theory of value*

The topic of measuring capital, particularly monumental ancient buildings constructed with

massive quantities of labor and probably only rudimentary capital implements, may lead some readers somewhere close to the labor theory of value from Classical economic theory of the late eighteenth and early nineteenth centuries. The tremendous diversity of capital implements being so patently obvious to the most casual of observers, to some scholars of the twentieth century the idea that there is in fact really only one ultimate factor of production, labor, and that all value can be expressed in terms of that numeraire, has contained considerable deepness. The idea of the labor theory of value is that, say, a bronze sickle blade can be represented in terms of the labor used in crafting it from previously smelted bronze, plus the labor used in smelting the bronze, plus the labor used to transport the ore from mine mouth to smelting site, plus the labor used in mining it; and that consequently the price of the sickle can be expressed in terms of this direct and indirect labor embodied in it.

Considerable research on the topic during the twentieth century has shown that a construct of total embodied labor in an array of goods will equal the competitive prices of those goods only if the interest rate is zero. The *relative* goods prices will equal relative total embodied labor inputs if relative prices are independent of the interest rate. Under this condition, this version of the labor theory of value is valid. However, this result comes with some implications: the price mark-up over embodied labor must be the same for each good; and the ratio of direct labor inputs to other inputs (indirect labor inputs in the form of capital) is the same across all industries. The fortuitousness of such a condition occurring has led neoclassical economics to the conclusion that the labor theory of value is wrong, and it has played no role in the development of the mainstream of economic thought in the twentieth century. In fact, to call the labor theory of goods pricing a theory of value overlooks the fact that it is strictly a cost theory – a labor theory of cost – because the utility of the goods produced – that is, demand – never enters the valuation process.

Let's return to the measurement of the capital value of an ancient monumental building,

constructed with a very small capital-labor ratio. What would be wrong with measuring the capital value by the labor used to construct it? If we simply estimated the labor hours (person-years) required to cut the stones, drag them to the site, put them in place, and smooth it all up, we would probably blur distinctions the ancients themselves made in what they were willing to pay for different kinds of labor: for example, rock-draggers, fine carvers, and architects. However, if we were willing to work with estimates of labor costs in terms of wage costs (or ration costs, if one so desired), and we added up the total wage cost of the building, we might not be so far off the mark. With a small enough capital-labor ratio, the cost of omitted tools *might* represent a small estimation error (their valuation relative to a unit of labor – hour, day, or year – might have been great enough to make the capital-labor ratio in value terms much larger than the same ratio in quantity terms). The cost of the stone with which the building was largely made (possibly mud brick as well) would have been only the labor cost of cutting and hauling it as long as no royalties had to be paid for those materials. Any external, environmental costs of the quarrying and hauling – denudation of hill slopes with attendant erosion, compacting of soil on the route from quarry to building site – would be beyond what it would pay to try to calculate, although at the time such costs could have been noticeable to the people directly affected. (Actually, compacting of soil on the hauling route might have conferred a positive externality if it helped compact a future road bed or performed unintended maintenance on an existing road bed.) In sum, labor costs of construction might be a reasonable, first-order approximation to the book value (but not necessarily the market value) of an ancient, monumental building.

8.2 Quasi-Rents

The marginal productivity model of factor pricing that we used in Chapters 1 and 2 depends on the ability of input quantities to be adjusted marginally, or “at the margin” – hence the name “marginal productivity theory.” By studying these

small (marginal) increments to the quantity of a factor employed we are able to attribute changes in output to the varying quantities of the input and hence are able to determine how much we could afford to pay for that particular quantity of that factor that is being varied. In those first two chapters we treated capital services as freely variable, but with a bit more realism we know that most capital takes quite concrete forms during the “short run”: buildings, tools, and so forth. They can’t be changed immediately into something else. This “fixed-factor” character is similar to the early nineteenth-century (particularly English) conceptualization of land, which was envisaged as being original, immutable, and in fixed supply, all of which we understand today to be mischaracterizations, even for the time at which the model was developed. Anyway, the term “rent” was developed to name the return to a factor that was in fixed supply, distinguishing it from the term “wage,” which was given to the other major category of input, which was readily observed to be in variable supply. In the late nineteenth century, the term “quasi-rent” was given to the return to a factor that was in fixed supply in some short run but was as variable in supply as any other factor in the long run.

The concept of quasi-rent is generally applied to the return to specific capital in short-run analyses. We offer an explanation of how quasi-rents operate in a series of diagrams. Figure 8.1 shows the marginal cost curve (MC) and the average variable cost curve (AVC) for some product. The price of the product is Op on the vertical axis (the

horizontal line pp can be thought of as either a perfectly flat demand curve or as simply a price line for this analysis). The firm to which these cost curves belong will produce to the point at which price equals marginal cost, which means that it will produce quantity Oq . At this combination of price and output, total revenue is the rectangle $Opcq$, but average variable cost is only area $Oabq$, leaving area $apcb$ as the payment to the fixed factor; this area is quasi-rent. Notice that if the price rose above p , the quasi-rent to the fixed factor would rise; notice also that it is the price that is determining the quasi-rent, not the quasi-rent that is determining the price. The sum of all variable costs (that is, the costs of all variable factors) and the quasi-rent equals the value of the total product, which is the same product exhaustion result we obtained from the sum of marginal products of variable factors in Chapter 1 (with constant returns to scale in production).

Now turn to Figure 8.2, which reproduces the cost curves of Figure 8.1 but adds another curve for average total costs. Area $apcb$ still represents the quasi-rent to the temporarily fixed factor, but now we can divide the quasi-rent into two parts: the opportunity cost (that is, what we have to pay to attract the fixed factor) and the temporarily pure profit to its short-run fixity. Let’s look at the effect of a sufficiently reduced price, say p' in Figure 8.3. This price still cuts the marginal cost curve above its intersection with the average variable cost curve, but it falls short of the average total cost curve. Output falls from

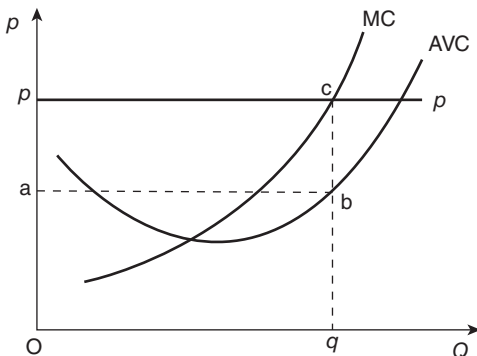


Figure 8.1 Quasi-rent.

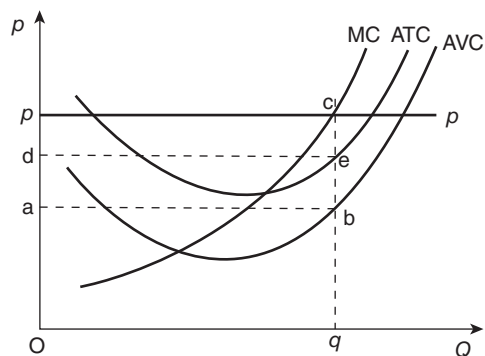


Figure 8.2 Components of quasi-rent.

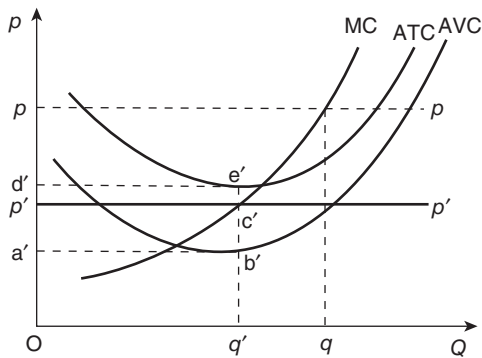


Figure 8.3 Effect of a price change on quasi-rent.

q to q' , which is reasonable, as the price facing a producer with a rising cost curve has fallen. Total quasi-rent is the area $a'p'c'b'$ (using the same letters as in Figure 8.2, but using "prime" notation to distinguish the different cases), but notice that this producer does not get area $d'p'c'e'$, which was pure profit in Figure 8.2. In this case, the total opportunity cost is area $a'd'e'b'$, which is greater than the quasi-rent by the rectangle $d'p'c'e'$, which in turn is an economic loss from employing this temporarily fixed factor in a use that gets only price p' per unit of output.

The "long run" is a period that is long enough to build new capital units and wear out old ones (in general, the time required for all factors to be freely variable). Over this period of time, the return on each piece of capital built must equal the current rate of return on capital. If a particular piece of equipment's quasi-rent is less than interest plus depreciation (quantities, not simply rates), it will not be replaced; if its quasi-rent exceeds that quantity, more pieces of equipment just like it will be built.

Finally, return for a moment to the present value formulations we have been using, in which we have taken the present discounted value of the rental or the vaguely defined "return" on capital in each of a number of future periods. Henceforward we will use the discounted stream of quasi-rents to obtain the price of a piece of capital.

8.3 Interest Rates

It is well known that interest rates were calculated and charged on various types of loans in Mesopotamia about as far back as we have found written records (for example, Postgate 1992, 169–170, 193–194; van de Mieroop 1992, 203–208). At times and places in the ancient Near East, including Egypt and the Aegean, for which such records do not exist, many scholars may be inclined to believe that interest rates did not exist. Implicit in such a belief is a theory of interest that depends on the existence of loans in a capital market for the existence of interest rates. In fact, interest is a much more general phenomenon and depends for its existence only on the linkage of economic actions over time, as we have begun to show implicitly in the present value formulation of a productive good (that is, a good that can be used to produce other goods). Consider the example of an ancient Egyptian shovel.⁹ It clearly was constructed with more than a single use in mind, possibly with a productive life of several years intended. Some of these items may actually have exchanged hands, either between maker and user or between users. Whatever was exchanged for the shovel could be termed its price, and if we (or they) figured hard enough, we could probably translate a price in terms of a commodity into a price in terms of a metal. Now, can we suppose that someone might have rented one of these tools out for a while? What the owner received for the use of the shovel over that period was its marginal product, its rental – its quasi-rent. If this shovel-lord (the shovel equivalent of a landlord) subsequently decided to part with his shovel for good – that is, sell it, what he would have been willing to accept for it would have been related to what the shovel could produce (for its former renter, and hence for its owner indirectly). The relationship between the per-period productivity of the shovel over its expected life time and the parting price acceptable to its owner would have been inversely proportional to the interest rate either implicit in the minds of the transactors

or explicit in some set of transactions where interest rates appeared transparently. This is a long-winded way of showing by example that the phenomenon we call an interest rate (our mental construct; also a mental construct with which ancient Mediterranean populations were familiar and comfortable) existed and was widespread throughout the ancient Mediterranean world, and helped regulate a wide range of activity that extended over lengthy time periods.

Nonetheless, some readers may still ask, “Where does this discount rate in the present value formula come from? The concept of a good providing services over several time periods is intuitively acceptable, and the notion of adding up that stream of anticipated production flows to get a parting price for such a good does no major violence to intuition either, but where does that interest rate come from and what *is* it?” These are reasonable questions, and they are the ones to which this section is devoted. Two partial answers are “the productivity of time” and “time preference.” You’ll note that one of these fits what we are coming to think of as the supply side of economic problems and the other fits the demand side. The concept of “the” interest rate, we will come to see, is actually quite restrictive; there is an “own-interest rate” for each good, making for a plethora of interest rates, many of which may move together over time – otherwise sharp transactors could make a bundle through arbitrage (the act of buying low and selling high in a fashion that eventually brings valuations of the same or substitutable goods into par with one another). We turn now to an exposition of own-interest rates.

To motivate the theoretical treatment of own-interest rates, we begin with a story – not necessarily a true story but one that certainly might have occurred and recurred. Imagine the relatively well-off family of a scribe in, say, Eighteenth Dynasty Memphis. The scribe has a regular relationship with a fisherman who brings a certain number of fish out of his catch to the scribe’s house every Wednesday morning. The fisherman selects the number of fish to add

up to a particular weight, so we consistently refer to a constant quantity of fish. Ignoring seasonal fluctuations in availability of fish, they long have agreed to a price of, say, two copper lumps totaling four grams as the price for these fish. They exchange fish and coppers each week. The fisherman has lots of other customers, and between them he is able to sell all the fish he can catch with the equipment he has at his disposal. Now, suppose that one week our scribe finds himself host to a swarm of unexpected visitors, and he believes that he should double the fish he has on hand to feed them. He gets word to the fisherman with whom he has the contract for the weekly fish delivery and asks him if he can bring both this Wednesday’s and next Wednesday’s loads of fish this Wednesday. The fisherman has a pretty tightly organized operation – he fishes about the maximum number of hours that the fish are biting, he routinely uses all the lines and hooks he has, and all his planned catch is allocated among his regular customers. What does he do? Just ignoring this request of his long-time customer could stand to lose him a customer, and besides, if he can pull it off, he might make a bit extra for his own family this week. But he has no extra resources of his own to expand his fish catch in order to move up the scribe’s next weekly order. What can he do to solve his problem? For one, he could see if he could rent some extra tackle from some other fisherman. For another, he could see if any other fishermen could be persuaded to part with some of their catches. Either way, it’s going to cost him, and by the time he gets finished scrambling around for a solution, he finds that the extra load of fish has cost him more than it would have had he been able to stick to his original schedule of delivery next week. For the regular load of fish, he can still charge the two copper lumps, but for the extra load, he has to charge, let’s say, two-and-a-half coppers (the half-copper denomination is no problem). This extra half-copper is the cost of moving up the delivery schedule on the next weekly “contract” amount.

Keeping in mind this cost of shifting planned deliveries from further in the future up to the present, let's consider the reverse direction of changing plans. Let's spin another story about the same scribe and fisherman. This time, the fisherman has had a hole knocked in the bottom of his boat, and he needs some time off from fishing to repair it, as well as a few extra coppers to acquire some materials he needs for the repairs. He has just about enough coppers stashed away to make the purchases he needs, but that'd cut things awfully close. He has a plan: he'll see if some of his customers would be willing to forego this week's fish delivery but still make their payment, getting, as it were, a week ahead in their contracted fish payments. He discusses this proposition with the scribe, who recognizes that he has some interest in seeing that his fisherman gets his boat fixed, but he also feels a little awkward about shelling out this week for something he won't get to eat until next week. So the scribe suggests a deal: he'll pay the fisherman one and three-quarters coppers this week for next week's fish. The fisherman thinks it over and figures that if he gets the one and three-quarters coppers now, he can find a way to use them to produce the equivalent of just about two coppers next week. So it's a deal.

These two stories have introduced several concepts that we will use in discussing own-interest rates. First, there is the scribe's intertemporal consumption plan regarding fish; without going into the details, the scribe and his family probably have future consumption plans regarding a large number of other goods as well. Second, there is the fisherman's intertemporal production plan, which, while satisfying the consumption plan of the scribe in question, also coordinates with the consumption plans of a number of other families. Third, the scribe's and fisherman's agreement to defer delivery of fish by a week at a price agreed on this week is a forward transaction. Fourth, the ratio of the copper-price of the one-week-forward delivery of fish to the copper price of a present delivery of fish is the own-discount factor for fish for a period of a week. This was the quantity of fish that had to be delivered now in exchange for the guaranteed delivery of the usual contract quantity in the following week.

Fifth, if we invert that own-discount factor (getting a number greater than 1) and subtract 1 from it, we have the own-interest rate for fish over the same period. We will use some tables below to lay out a general intertemporal price pattern and will show some specific formulas relating prices for different dates to own-interest rates.

Looking across each row of Table 8.1, we have an intertemporal equilibrium set of prices for goods 1 through n , from time 1 (the present) to time T . Our stories above offer some worm's-eye explanation of how these prices originate. Now, we need to find a numeraire in which to measure these prices. While there are many possible such numeraires, we can arbitrarily take good 1 in time period 1 as the numeraire by multiplying every p_{ij} in the table by $1/p_{11}$. This gives us Table 8.2,

Table 8.1 Intertemporal equilibrium prices.

Goods	Time periods			
	1	2	...	T
1	p_{11}	p_{12}	...	p_{1T}
2	p_{21}	p_{22}	...	p_{2T}
3	p_{31}	p_{32}	...	p_{3T}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
n	p_{n1}	p_{n2}	.	p_{nT}

Source: Adapted from Bliss (1975, 50, Table 3.1), by permission of Elsevier/North-Holland.

Table 8.2 Present-value relative prices.

Goods	Time periods			
	1	2	...	T
1	1	p_{12}	...	p_{1T}
2	p_{21}	p_{22}	...	p_{2T}
3	p_{31}	p_{32}	...	p_{3T}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
n	p_{n1}	p_{n2}	.	p_{nT}

Source: Adapted from Bliss (1975, 51, Table 3.2), by permission of Elsevier/North-Holland.

which looks just like Table 8.1 except that p_{11} in the top row and left-most column has been replaced by a 1. Each p_{ij} in Table 8.2 is the original p_{ij} of Table 8.1 divided by p_{11} . These prices can be called present-value relative prices.

Suppose that one of the people in this economy holds one unit of good 1 in period 1. If he wanted to exchange that good for some amount of the same good in a later period, the most he could obtain with his first-period holding is some quantity q_{1t} ; the price of good 1 in period t is p_{1t} , so we can calculate that the amount of period-1 stock he holds, 1, can command $p_{1t}q_{1t}$, so $1 = p_{1t}q_{1t}$, and consequently the quantity of good 1 he can obtain in week t by trading on his period-1 holding is $q_{1t} = 1/p_{1t}$. In general, the present-value prices in Table 8.2 are the inverses of the quantity of good i for delivery in period t that one unit of the numeraire good (good 1 in time 1) will command. Pursuing this type of trading a bit further, the additional amount of good 1 that this person can get in week t by offering one unit of the numeraire good is $(1/p_{1t}) - 1$, which is the definition of the own-rate of interest of good 1 between times 1 and t , $r_{1(1,t)}$, where the parenthetical portion of the subscripting identifies the beginning and ending periods over which the own-rate is calculated. More generally, the own-interest rate for any good i , across any pair of periods t and $t+k$, is $r_{i(t,t+k)} = (p_{it}/p_{i,t+k}) - 1$. The own-rates of interest on different goods, over the same time period, are generally not equal. They will be the same only if their relative prices are unchanged over the period.

Do own-interest rates have to be positive? No. An own-interest rate will be positive as long as the own discount factor, $p_{i,t+k}/p_{it}$, is less than 1; if the discount factor is greater than 1, the associated own-interest rate will be negative, but it cannot fall below 1.0 itself. The own discount factor usually will be less than 1, provided that the productive processes are able to expand the quantity of goods available over time. In such cases, present-value prices have to decline with time to eliminate pure profits – arbitrage will accomplish this.

An equilibrium intertemporal price system contains within itself an entire structure of own-rates of interest – separate own-rates for each good and each pair of time periods. We can convert this structure of prices of Tables 8.2 into a combination of own-discount factors and own-rates of interest for good 1. Rearrange the expression for the own-interest rate of good 1 in the previous paragraph to relate good 1's period-1 price to its price in subsequent periods: $p_{11} = (1 + r_{1(1,t)})p_{1t}$. Now, use this expression for the price of good 1 to create Table 8.2's relative prices in the format shown in Table 8.3, in which each price is expressed as the product of a relative price in a given time period and the discount factor for good 1 for the relevant pair of periods.

Now, let's revisit the simplifying assumption that complete forward markets exist to reveal this intertemporal price structure. As a matter of fact, there are few forward markets even in today's industrialized economies, and those that

Table 8.3 Discounted future relative prices.

Goods	Time periods			
	1	2	...	T
1	1	$1/[1 + r_{1(1,2)}]$	...	$1/[1 + r_{1(1,T)}]$
2	p_{21}/p_{11}	$(p_{22}/p_{12})/[1 + r_{1(1,2)}]$	...	$(p_{2T}/p_{1T})/[1 + r_{1(1,T)}]$
3	p_{31}/p_{11}	$(p_{32}/p_{12})/[1 + r_{1(1,2)}]$	...	$(p_{3T}/p_{1T})/[1 + r_{1(1,T)}]$
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
n	p_{n1}/p_{11}	$(p_{n2}/p_{12})/[1 + r_{1(1,2)}]$	.	$(p_{nT}/p_{1T})/[1 + r_{1(1,T)}]$

Source: Adapted from Bliss (1975, 54, Table 3.3), by permission of Elsevier/North-Holland.

do exist yield prices for only a few dates. In the absence of such a generator of “complete information,” economic agents will make guesses for the forward prices. While this incompleteness of information can prevent the attainment of a complete, intertemporal equilibrium that these tables make look so easy, these present-value prices, arrived at by guesswork or an entire information industry, remain the forces that direct economic decisions.

8.4 The Theory of Capital

How much capital would a particular economy be willing to hold, wanting neither more nor less? If this economy finds itself with a capital stock either larger or smaller than this desired amount, how rapidly would it change its stock? Answers to these questions determine the stationary equilibrium stock of capital, the equilibrium rates of saving and investment, and the interest rate. The interest rate (or the array of interest rates, as we observed in section 8.3) affects the structure of production, the division of total output between consumption and investment, the time pattern of consumption of individuals, and the forms in which assets are held. It's an influential variable.

We can look at capital theory from two perspectives. We can think in terms of the purchase and sale of permanent income streams, with the demanders of permanent income streams being the owners of capital and the suppliers being producers of those permanent income streams. Alternatively, we can view the production units as demanders of capital and savers as the suppliers of capital. The two views are equivalent and complementary. We will use both of them in the exposition below.

8.4.1 *Present and future consumption, investment, and capital accumulation*

We begin with a model in which the capital good is the same as the consumption good: if not consumed in the present period, it simply grows into a larger quantity of the consumption

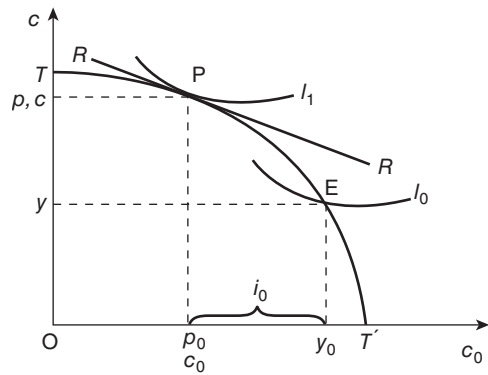


Figure 8.4 Relationship between present and future consumption.

good in the subsequent period. After making a few points with this very simple model, we will switch to a model that uses a capital good that is distinct from the consumption good. Figure 8.4 shows the combinations of current consumption, c_0 , and levelized, perpetual consumption, c , between which an individual may choose, subject to her initial endowment of productive goods and labor.¹⁰ Our individual begins with an endowment of current and future consumption defined by point E. Through point E, we have drawn the set of transformation possibilities TT' that govern how this individual might convert the current and future consumption represented by point E into other current and future consumption combinations. Curve TT' is the familiar transformation curve, anchored by endowment point E. If our individual chose to remain at the endowment point, she would obtain current income y_0 and a perpetual future income of y . But given her preference system, represented by indifference curves I_0 and I_1 , she could find a preferable combination of current and future consumption at point P, where indifference curve I_1 is tangent to the transformation frontier. This tangency yields an interest rate equivalent to the absolute value of the slope of line RR' and gives current production and consumption of $p_0 = c_0$ in exchange for levelized perpetual production and consumption of $p = c$. Note that p_0 and c_0 are less than what the individual could have produced and consumed at the endowment point, y_0 , but that p and c

are greater than the perpetual future income that could have been had in combination with current income y_0 . We can look at the difference between y_0 and (p_0, c_0) in two ways. From the perspective of production, $y_0 - p_0 = i_0$, where i_0 is current-period investment. By setting aside some current productive power, our individual is able to produce more in the future. Looked at from the perspective of our individual as a consumer, $y_0 - c_0 = i_0$, which says that by foregoing some current consumption, our individual can consume a larger slice off the larger capital stock next period. The difference $p - y$ (or $p - c$) on the vertical axis is the return to the investment i_0 on the horizontal axis. (Note that $y_0 - p_0$ and $y - p$ must have different signs.) The distance $y_0 - p_0$ need not be positive, as we have drawn it in Figure 8.4; a positive distance characterizes investment, but a negative distance (that is, p_0 located to the right of y_0) would characterize disinvestment, a situation in which the existing capital stock was larger than desired.

The intersection of the transformation curve TT' with the horizontal axis is the greatest amount of consumption that the individual could obtain if she were to consume both her maximum currently produced income, y_0 , and the capital with which she entered the current period, K_0 . Of course, if she does such a thing, she will consume nothing in any future period. If she invests the amount i_0 this period, she will enter the next period with an amount of capital equal to $K_0 + i_0$.¹¹ The distance on the horizontal axis between T' and y_0 is the current capital stock, K_0 – that is, the amount with which our individual entered the current period; she departs the current period (or, equivalently, enters the next period) with $K_0 + i_0 = K_1$.

Figure 8.5 presents the same information as Figure 8.4 but from the perspective of the production of the consumption good with varying amounts of capital and a constant labor input. The horizontal axis measures the total capital stock, and the vertical axis shows the magnitude of the levelized, perpetual production stream p . Curve TP shows the total future product attainable with particular sizes of capital stock. As we move from the origin out to the right on the horizontal axis, we proceed through time

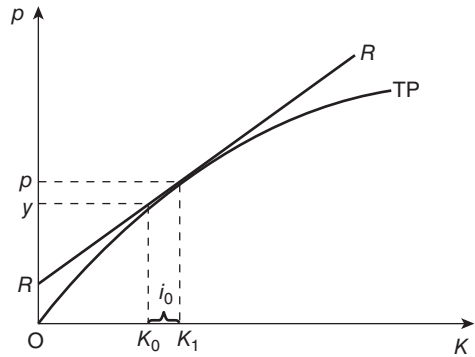


Figure 8.5 Marginal product of capital.

as well as increasing the quantity of capital. The TP curve shows decreasing marginal returns because the increasing quantity of capital is being employed with a constant quantity of labor. Line RR is, again, the interest rate associated with point P in Figure 8.4. Vertical distance $p - y$ is the return from the incremental capital stock i_0 (or, alternatively, the return to the investment). Think back to Chapter 1 and you will recall that line RR, being tangent to the total product curve, is the marginal product of capital; it is also the interest rate. Consequently, when the output (the consumption good) is exactly the same material as the capital good, and there are no diminishing returns in converting deferred consumption (investment) into new capital goods, the interest rate is equal to the marginal product of capital. This is the only case in which this simple equation is true.

This model may appear quite simple. To give a peek at just how complicated this simple model really is, let's count variables. The model determines the values of the following variables: current and annualized future consumption (c_0 and c), current and annualized future production (p_0 and p), the current endowment production possibility (y_0), the next period's beginning capital stock (K_1 , equivalent to the endowed capital stock, K_0 , plus this period's investment, i_0), the equilibrium rate of tradeoff between current and levelized future consumption ($\Delta c / \Delta c_0$),¹² the marginal product of capital ($\Delta p / \Delta K$), and the interest rate (or the inverse of the price of a unit of the capital good). Each of the first eight of these

variables must be determined for each individual in the economy; the interest rate is common to all. A specific relationship (an equation) is associated with each variable. This model has $8N + 1$ variables and equations, where N is the number of individuals in the economy.

You will have noticed that, while there is an equilibrium in each time period in this model, it is difficult – in fact impossible – to say that the capital stock is in equilibrium, because we appear to have a positive investment in each period (actually, we could have negative investment, but either way it makes no difference: the size of the capital stock keeps changing, and the capital-labor ratio with it). We turn to this issue next, but we do that in two stages. First, we will use diagrams like Figures 8.4 and 8.5 to follow the accumulation of capital from some initial endowment point to a stationary state – the level of capital at which the stock is at the desired level. Figure 8.6 follows the accumulation of capital by a representative individual, beginning in some time period zero with endowment E_0 , for which there exists transformation frontier T_0 . Given this individual's preference system, represented by indifference curve I_0 , the choice of production points is P_0 , which yields an investment i_0 which is not shown in the diagram. The period-zero interest rate is represented by the slope of line R_1 . With the passage of time, production point P_0 becomes the new endowment

E_1 , along the 45° line, by conferring on time 1 current-consumption claims $p = y_1$ and a new capital stock enlarged by the previous period's investment. The individual can reach a higher indifference level, I_1 , at new production point P_1 . With the investment of time period 1, production point P_1 becomes endowment point E_2 with the addition of the investment i_1 . The interest rate in time 1, with the larger capital stock, is lower than it was in time period zero, represented by a somewhat flatter slope of line R_1 (compared to R_0). A similar process of accumulation (growth of the capital stock) continues until the desired production point is identical to the endowment point, at which time we have reached the stationary state, in which the actual capital stock equals the desired stock, and investment is zero. This is represented by $P_s = E_s$ on indifference curve I_s (transformation frontier T_s is not drawn). At this size of capital stock, the interest rate reaches its minimum, with line R_s , flatter than any of the previous R_i lines. The accumulation line AA traces out the series of production points on the way to the stationary state. Figure 8.7 shows the progress of capital accumulation and the equilibrium interest rate along the total future product curve.

Before departing this model, we offer two alternative depictions to leave the reader with the understanding that Figure 8.4, and the other figures deriving from it, could contain other

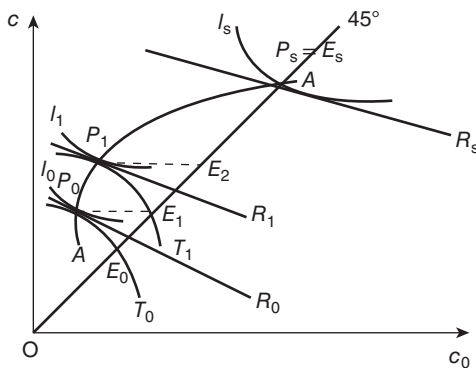


Figure 8.6 Accumulation of capital by an individual.

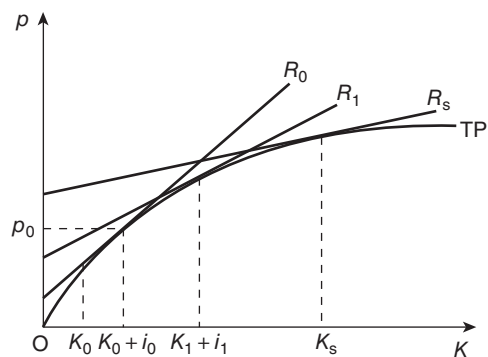


Figure 8.7 Relationship between capital accumulation and the interest rate.

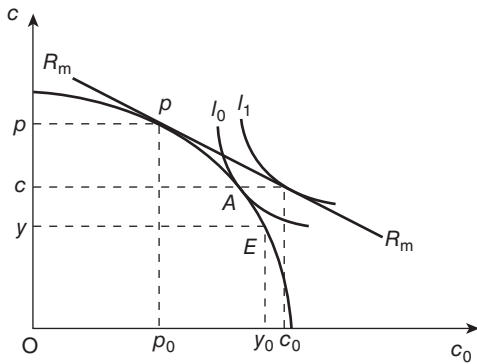


Figure 8.8 Intertemporal consumption choices, with and without market opportunities.

configurations of production and consumption than those drawn so far. Figure 8.8 shows an individual facing a given interest rate, represented by line R_m . If this individual were to be kept from transacting in the market indicated by R_m , beginning with endowment E , he would produce at point A (for which the corresponding p , p_0 , y , and y_0 are not drawn), reaching indifference level I_0 . Once he is allowed to transact in the market, production and consumption, both present and future levelized, diverge, and he can reach the higher indifference level I_1 . Current consumption can rise, but levelized future consumption falls.

Figure 8.9 complicates the capital good by making the consumption good distinct from

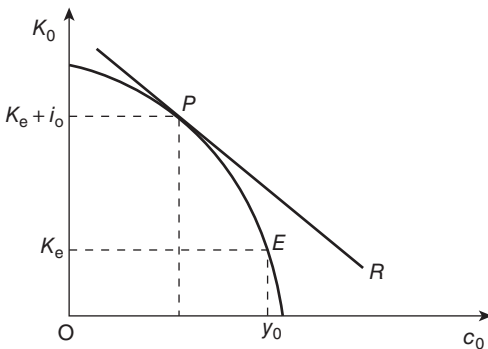


Figure 8.9 Tradeoff between current consumption and capital accumulation.

the capital good. The direct tradeoff the consumer faces is between current consumption and the level of capital stock with which he will produce. Point E is the endowment point, and line R is the interest rate facing the consumer. The diagram does not contain an indifference curve because the capital good, K_0 , is strictly an intermediate good, while the consumer has an indifference relation only between current and future consumption. Whereas the interest rate in the simpler model was $\Delta p / \Delta K_0$, the marginal (future) product of capital, the relationship between the interest rate and the marginal product of capital is less direct. With the specification of the capital good as distinct from the consumption good, the interest rate is equal to the product of the marginal product of capital and the marginal product (in terms of capital) of investing: $(\Delta p / \Delta K_0) \cdot (\Delta K_0 / \Delta i_0)$, or letting the terms cancel, $\Delta p / \Delta i_0$, which can be called the marginal efficiency of investment (called a marginal efficiency rather than a marginal product because it is a pure number: its numerator and denominator are measured in the same units and hence cancel).

8.4.2 Demand for and supply of capital: flows and stocks

We will present the supply and demand curves for capital from two alternative perspectives. First, we will use the construct of the permanent income stream as the object that is demanded. We do this to see how saving, investment, and consumption are related in the determination of the equilibrium capital stock. Then we will focus directly on stocks of capital. Be careful to note that we measure $1/i$ ("one over the interest rate") on the vertical axis of the permanent income stream diagrams and i (just the interest rate) when we work with the capital stocks. Figure 8.10 shows a fixed supply of permanent income streams (suppose the permanent income streams are measured in either shekels or sacks of barley; it doesn't matter) as the vertical line labeled S ; the demand curve for permanent income streams is D . The price of $1/i^*$ will exactly equate this society's demand for permanent income streams

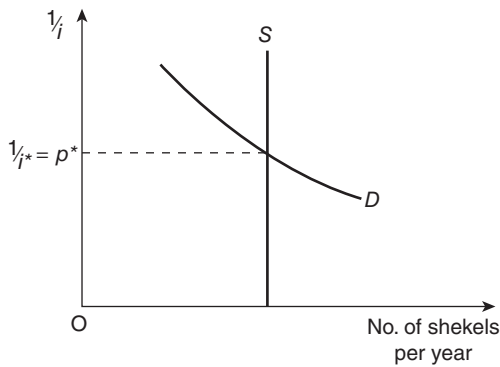


Figure 8.10 The stock and price of capital.

with the fixed supply. The supply curve refers to a stock of permanent income streams, as does the demand curve. The supply and demand curves we have worked with in previous chapters have been for flows of things – so many units per month, year, or whatever time period we want to use. Figure 8.10 shows, for a given stock of capital, how much people will be willing to pay per unit of it.

Now, in Figure 8.11 we introduce the possibility of adding to or subtracting from the stock of capital. Retaining P^* as the stock equilibrium price, let's suppose that the price for capital is P^+ . At this price, more permanent income streams are offered for sale than people are willing to buy. How does this discrepancy get equilibrated? The extra income streams can be converted into

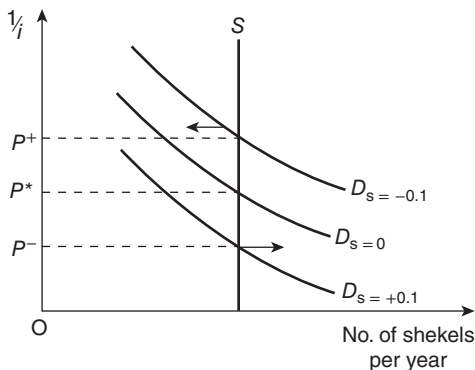


Figure 8.11 Altering the stock of capital.

consumption by dissaving. Dissaving raises the level of current consumption relative to the supply of permanent income streams, which causes consumers to attach a lower value to further additions to current consumption, relative to their valuation of the income streams. Larger rates of dissaving have a stronger effect along these lines. Thus, there is some level of dissaving that will raise the price that people are willing to pay for a permanent income stream to P^+ . We represent this with the curve in Figure 8.11 labeled $D_{s=-0.1}$, representing the demand for permanent income streams when the dissaving rate is equal to (an arbitrarily chosen example of) 10% of income. But this position can't be maintained, since the dissaving reduces the size of the capital stock, as indicated by the left-point arrow near the intersection of S and $D_{s=-0.1}$. Correspondingly, if the price of an income stream had been P^- , people would have wanted more income streams than were available, they would have saved part of their income (represented by $D_{s=+0.1}$), and the supply of capital would move to the right. Only curve $D_{s=0}$ is consistent with a long-run stationary equilibrium – that is, the situation in which society has exactly the capital stock it wants.

Figure 8.12 shows a supply curve for the stock of permanent income streams as increasing in the price of a unit of the stream, which is the inverse of the interest rate. Along this supply curve, investment is zero; in this sense it is comparable

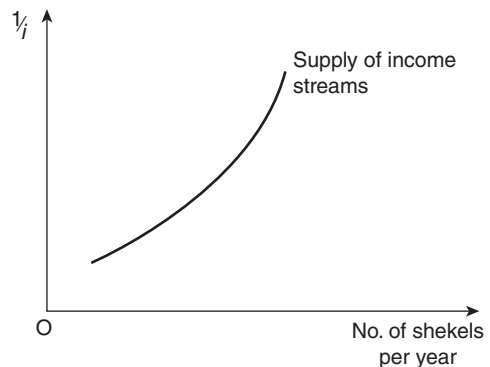


Figure 8.12 Supply curve of permanent income streams.

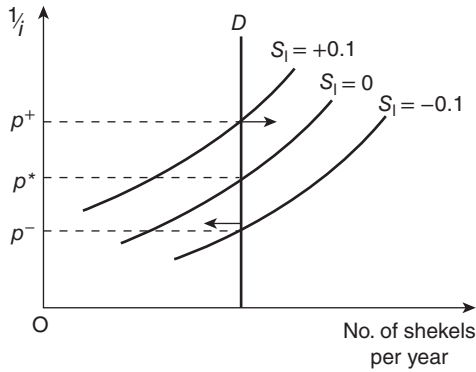


Figure 8.13 Investment and disinvestment.

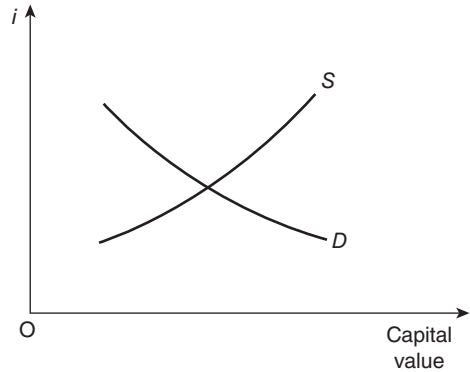


Figure 8.14 Supply and demand curves of the stock of capital.

to the demand curve shown in Figure 8.10 (or curve $D_{s=0}$ in Figure 8.11). In Figure 8.13, the zero-investment, stock supply curve is accompanied by momentary stock supply curves representing investment when the price is above the stationary equilibrium price ($S_{I=+0.1}$) and disinvestment when it is below ($S_{I=-0.1}$). As with the zero-saving demand curve, the zero-investment supply curve is different from the supply curves with either positive or negative investment.

Now, we will look at this problem from the other perspective. The demand for capital will be the demand coming from producers who use capital; in the discussion surrounding the previous four diagrams, this was equivalent to the supply of permanent income streams. And now the supply of capital will be what savers offer up by not consuming; this was the demand for permanent income streams in the previous analysis. In Figure 8.14, S and D represent these reformulated, stock supply and demand curves for capital. However, with this formulation of the value of capital, the value of any particular, physical capital stock will vary with the rate of interest, and in Figure 8.15, curve C is a rectangular hyperbola representing the value of a constant stock of capital as the interest rate changes. Since the demand curve D is the demand from producers, at an interest rate as high as $i_{\max}$, they would not want to have more capital since they would just have to pay more interest on it. At interest rates this high, however, savers would be delighted to

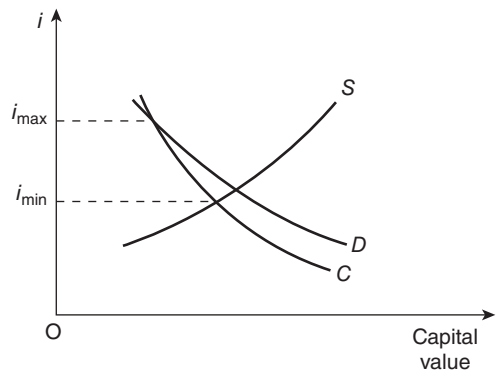


Figure 8.15 Influence of the interest rate on a given stock of capital.

lend more. The lower the interest rate, the further the demand curve is above the C curve. At an interest rate as low as $i_{\min}$, suppliers of capital (savers) have no incentive to lend more, although the capital-demanding firms would be glad to use more. So the interest rate in any temporary equilibrium of saving and investment on the way to the stationary equilibrium must be between $i_{\max}$ and $i_{\min}$.

Figure 8.16 puts together both stationary- and temporary-equilibrium, stock supply and demand curves for capital values. At higher interest rates, savers – the suppliers of capital to producers – would like to invest larger proportions of their income, represented by the upward

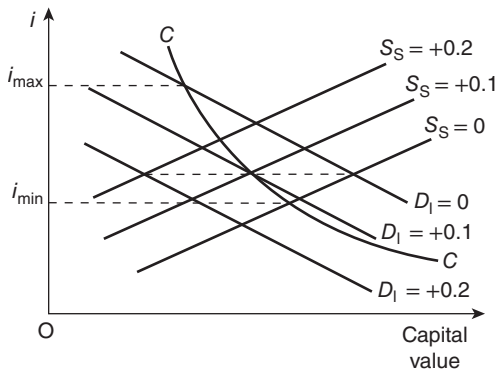


Figure 8.16 Stationary and temporary equilibria in the demand for and supply of capital stock.

progression of stock supply curves $S_{s=0}$, $S_{s=+0.1}$, and $S_{s=+0.2}$, with the subscripts indicating the proportions of income being saved along each curve. Demanders would want to invest more at lower interest rates, represented by the downward progression of the demand curves $D_{I=0}$, $D_{I=+0.1}$, and $D_{I=+0.2}$. The intersection of demand curve $D_{I=+0.2}$ and supply curve $S_{s=+0.2}$ is a temporary saving-investment equilibrium, and similarly with the intersection of $D_{I=+0.1}$ and $S_{s=+0.1}$. The intersection of the zero-saving supply curve, $S_{s=0}$, and the zero-investment demand curve, $D_{I=0}$, is

the stationary equilibrium yielding the long-run equilibrium stock of capital value.

An alternative view of the flows and stocks is offered in Figures 8.17(a) and 8.17(b). Figure 8.17(a) shows the flows of current investment and saving (single-period *changes* in capital stock), and Figure 8.17(b) shows total capital stocks, so the horizontal scales are vastly different. Accordingly, the stationary stock demand and supply curves in Figure 8.17(a) are nearly horizontal relative to the single-period investment (I) and saving (Σ) curves. In Figure 8.17(a), if the interest rate goes high enough, investment (flow demand from producers using capital) will go negative; if it goes low enough, saving from consumers will go negative. At the vertical axis in Figure 8.17(a), we are effectively at the previous period's capital stock: the flow saving curve intersects the zero-saving, long-run stock supply curve. At the capital stock so defined, if we trace the interest rate over to Figure 8.17(b), we are at capital stock K_i , which is consistent with a very low interest rate on the zero-saving supply curve and a very high interest rate on the zero-investment demand curve. The current-period addition to the capital stock, denoted $\Sigma = I$ in Figure 8.17(a), moves the total capital stock to K_{i+1} in Figure 8.17(b), that addition being consistent with the interest

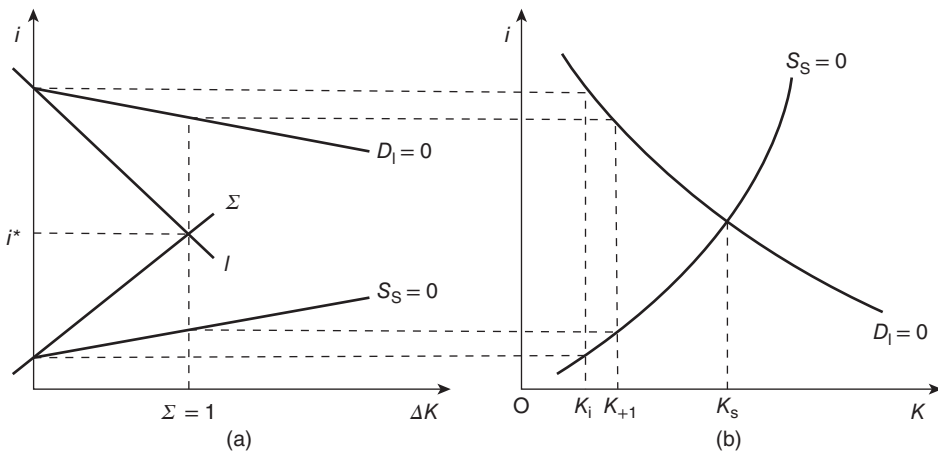


Figure 8.17 (a) Capital accumulation: flows of current investment and saving. (b) Capital accumulation: resulting stock supply and demand curves.

rate-capital stock combinations denoted by the dashed lines from the $\Sigma = I$ intersection to the $D_{I=0}$ and $S_{s=0}$ curves in Figure 8.17(a). Over time, this economy will move toward the long-run, stationary equilibrium capital stock K_s in Figure 8.17(b).

Why have I subjected you to this discussion? Several important issues emerged in what we've just been through. First, the short-run relationship between the interest rate on the one hand and investment and saving on the other is different from the long-run relationship between the stock demand for and supply of capital. In the long-run equilibrium, saving and investment (ignoring the replacement of depreciated capital) will be zero at what will probably look pretty much like a "normal" value of the interest rate – that is, neither shockingly low (zero saving) nor high (zero investment). Second, we get some as-yet dim insights into the relationship between how rapidly an economy can approach its desired capital stock from any given stock size and how costly that movement may be. Third, we have the basis for understanding how any given size of capital stock might be consistent with either high or low interest rates, a subject to which we turn in the following subsection. Fourth, the connection between current consumption and capital formation contains a number of subtle valuation relationships that help determine the equilibrium of consumption and saving. Particularly, changes in the *rate* of consumption alter the price of capital in the same direction: more current consumption, higher capital price, more attractive investment (less attractive savings opportunities). Stories about the accumulation of capital in antiquity need to take account of these relationships.

8.4.3 Capital richness and interest rates

It may be tempting to associate lots of capital ("capital richness") with low interest rates. This subsection explores a number of possibilities in this association. First recall that a large quantity of some good (produced capital goods, for instance) typically depresses its price relative to the prices of other, less abundant goods, and the same phenomenon should apply to capital as

well. But a lower capital goods price is associated with a higher interest rate, giving us our first hint that there is more than may meet the eye to the temptation described in the opening of this paragraph. Rather than developing further tools, this subsection will discuss some results of what could be called sensitivity analysis of a relatively abstract model of capital accumulation (specifically, semi-stationary growth, or expansion of an economy in size, with a constant structure of factor and price proportions) originally developed and analyzed by Bliss (1975, Chapter 4, esp. 78–85). As I have pointed out on several occasions, particularly in this chapter, no economist believes that any real economy looks like the abstract model under analysis, but the severe abstractions have proven necessary to gain basic understandings and do prove useful in highlighting how particular relationships operate, as well as frequently suggesting how the omission of more realistic details would affect the results.

The technique is to compare the growth paths of alternative model economies that differ only in specified attributes. These economies use produced capital goods and nonproduced inputs (such as labor and land) to produce consumption goods and, of course, more capital goods. The model is one of general equilibrium, in which many of the simple, partial-equilibrium results such as a larger quantity of a good being available depressing its price, are contingent on the circumstances of many other inputs and outputs.

Let's begin by comparing the growth paths of two economies that differ only in their ratios of capital to nonproduced productive services. The growth path of the economy with the higher capital ratio cannot have lower total earnings of all factors, when earnings are measured in terms of the consumption good. Where capital is relatively plentiful, the other group of inputs will be relatively more highly valued. When these growing economies reach their stationary states (that is, when growth is not occurring), the economy with the higher ratio of capital to nonproduced inputs cannot have a higher rental rate of the capital good in terms of the consumption good. However, on a path of growth, it is possible for the economy with the higher level

of capital to have the higher capital rental price as well. The intuition that capital services should command a lower rental when capital is more abundant allows only for the influence of the cost of the capital service on the demand for it as an input, excluding the necessary incentives for the production of lots of capital. In the semi-stationary growth under examination here, the more capital-rich path also has to produce more capital, as well as use more of it, so the price of the capital good must be high enough to make it attractive and possible to do so. There are two forces in this result, pulling in opposite directions. First, a large *production* of the capital good, relative to the consumption good, makes for a high capital good price (working up the supply curve), which in turn puts upward pressure on the rental rate. Second, the *use* of a large quantity of capital services relative to, say, labor depresses the capital rental rate. In the stationary state, there is zero net production of the capital good, which eliminates the first influence, so the intuition described above really refers to a situation of no growth.

Second, a capital-rich growth path may have a higher interest rate than a capital-scarce path. In a situation of more abundant capital, the cost of capital is lower, that cost being the rental on the capital good. The interest rate is not the cost of using capital, and even a stationary, more capital-rich economy can have a higher interest rate than would exist on a less capital-rich growth path. But, a growth path with a higher ratio of capital to nonproduced factor services cannot have a higher interest rate than a growth path of a more capital-scarce economy. As long as there is more than one kind of output (that is, in any situation more complicated than Knight's Crusonia Plant model), a capital-rich growth path can have a higher interest rate than a capital-scarce growth path.

The facts that (i) capital is both an input and an output and (ii) its contributions to output are defined by intertemporal stock-flow relationships make capital a particularly intricate subject of study. General equilibrium models generally are required for the study of capital, and such settings permit counterintuitive results to emerge frequently. Having said this latter, we should hasten

to note that the simple positing of a counterintuitive relationship between economic variables rather than deriving them from a model is not a useful procedure.

8.5 Use of Capital by Firms

In section 8.4, we treated capital pretty abstractly. In this section we turn to the analysis of capital at the level of the individual producer, which, while remaining somewhat abstract in some points, at least permits us to address issues that are more closely related to directly observable phenomena, such as the number of buildings or total floor space used by a manufacturer or distributor. We focus on the three related issues of investment, maintenance, and the decision to retire particular units of capital equipment.

8.5.1 Investment

Investment is the addition of new (or second-hand) capital to an existing stock of capital. The important question facing the investor is how much capital (new productive capacity) to add in any particular period, but this issue itself implies several other decisions – how quickly the producer should approach a target capital stock; when investments should be made; and how to respond to new developments that emerge during an investment planning horizon.

“How much” questions are best thought of in terms of maximizing something, but the owner or manager of a firm could choose one of a number of targets to maximize: the average internal rate of return (irr), the marginal irr, the ratio of firm present value to present costs, or the net present value of the firm (present value of revenues minus present value of costs). The average irr is the interest rate that would make the total capital investment just break even. The marginal irr is the rate that, if used to discount the marginal revenue from an additional unit of investment, makes the present value of that revenue equal to exactly one. If the marginal irr equals the interest rate, this criterion is equivalent to maximizing net present value. The best criterion to use for different kinds of capital projects (another

term for an investment) has been the subject of considerable research in the past half century, with the result that, while net present value is considered generally the best maximization target, even it needs to be used with care because the time stream of revenues and costs can cause problems.¹³ I say this not because I expect readers to be concerned with “picking the best projects” themselves but because it’s not obvious which criterion ancient investors might have used in any particular case. There were plenty of alternatives, those alternatives have been subject to theoretical investigation quite recently, and there long have been differences of opinion about maximization targets. Additionally, accounting concepts are needed to implement any of these maximization goals, and accounting has developed considerably over the centuries as well. We are able to offer limited guidance on the precise specifics of what ancient investors would have attempted to do, but more on the consequences of whatever they did do.

Let’s talk about “policies” of producers for a minute. It is common to think of government plans when using the term “policy,” but the term can be used to refer to the set of plans of any decision maker. In this latter sense, a producer will have an output plan, an input plan, a maintenance plan, a replacement plan, an expansion plan, and probably several others as well. Each of these plans is interdependent with the others, and all together, they will comprise a set of plans that the producer expects to be optimal, in the sense that if he thought he could do better with the resources at his disposal and the circumstances in which he expects to operate, he would. In other words, we don’t expect people, now or in the past, to intentionally and systematically leave productive possibilities “on the table.” We will refer to policies and plans of investors in the rest of this section.

Before diving into the theory of investment we need to talk for a moment about the role of expectations in investment. Investments are forward-looking actions. They have most of their consequences in the future – all of their benefits and many of their costs, although most current decisions about the future will commit the decision maker to some current actions that

are irreversible. As we discussed in Chapter 7, economic expectations have to have some connection to something else that itself is known. Consequently, investors use their current knowledge to make their best estimates of what various conditions will be like in the future. They will be concerned more with the nearer future than the farther future for two reasons: nearer events are generally more predictable than more distant events, and, even for perfectly known events, nearer ones need be discounted less than later ones and consequently weigh more heavily in the assessment of benefits and costs. When contemplating an investment policy – either a continuous stream of investments over some period of years or a single investment at a single time – an investor implicitly or explicitly (explicitly being better) relies on expected time paths of input prices (of both the investment good and coordinating inputs), output prices (the time path of demand for the firm’s output), technological progress, and government policies (taxes, trade policies, and so forth) at the very least. Events in all of these categories will affect future revenues and future costs which, of course, are the raw material of net present values and any other investment criterion. An investor might reasonably expect some prices to remain constant over his planning horizon. Others might be expected to rise or fall continually, or be subject to cycles of more or less predictability. As we noted in Chapter 7, unfolding events can alter the entire, remaining time path of expectations right in the middle of an old (previous) projection period. Expectations of future events, if held with considerable confidence, can alter anticipatory behavior even prior to the occurrence of the event. For instance, if a tax change is either announced or strongly expected, it will affect investment behavior as soon as the new expectation is formed.

Having prepared ourselves with a few preliminary concepts, we turn now to the theory of investment, which we will characterize primarily as the demand for investment. In Chapter 1 we developed the concept of the demand for the services of capital, and we did that in a static context, that is, without explicit concern for the temporal dimension of production. There is nothing wrong

with that; static analysis is a legitimate procedure for studying long-run tendencies. However, while static analysis yields a meaningful concept of the demand for capital services, dynamic, or intertemporal, analysis is required to produce a demand for investment. The entire future time path of a firm's production plan is required to determine investment demand. The problem is set up as a constrained maximization problem, just as has characterized the solution of our optimization problems in the static setting. The firm owner or manager (or whoever is charged with making the production decisions, including the investment decision) maximizes the net present value of the firm subject to two constraints, one characterizing the production technology (that is, the production function), the other defining the firm's capital accumulation: maximize $W = \sum_t [p_t Q_t - w_t L_t - q_t I_t](1+i)^{-t}$, subject to $Q = f(K, L)$ and $\dot{K}_t = I - \delta K_t$. In this expression, the subscript t refers to the time period; p_t is the output price expected at each time t (this is an example where an entire time path of a variable must be projected either implicitly or explicitly by the firm manager); Q_t is the time path of planned output; w_t is the labor cost at each time; q_t is the price of the investment good at each time; I_t is planned investment at each time t ; $\dot{K}_t$ (read it as "K-dot") is the time-rate of change of the capital stock, $(\Delta K/K)/\Delta t$, at each time; and δ is the depreciation rate, that is, the rate at which capital in use "decays."

The full solution of this maximization problem determines the time paths of output, inputs, investment, and the cost of capital: Q_t , K_t , L_t , I_t , and c_t . As one of the first-order conditions of this maximization problem, we set the marginal product of capital, $\Delta Q/\Delta K$, equal to the marginal cost of capital, c_t/p_t , one of the usual conditions for profit maximization (in this case, maximization of the firm's value). In the expression for marginal cost, c_t is the user cost of capital, which after a number of further manipulations, is found to equal (omitting the time subscripts) $q(I + \delta) - \dot{q}$, where the dot over the q at the end of the expression has the same meaning as it did over the K , a time-rate of change, but in the capital price in this case. This condition tells us that the cost of capital has three components: the marginal

product, or rental rate, of capital, q_t ; depreciation on the capital attributable to its use, $q\delta$; and any decrease in the price of the capital good, q . Recall from section 8.1.5, in the discussion of the age composition of a capital stock, we noted that an increase in the price of capital in a firm's capital stock reduces the user cost of capital.

As part of the solution to this problem, the investment constraint, $\dot{K}_t = I_t - \delta K_t$, is rearranged to give the investment rule $I_t = \dot{K}_t + \delta K_t$, which warrants some discussion. First, this equation is the demand for investment. Second, this result divides investment into replacement of depreciated capital, δK , and expansion investment that results in a change in the capital stock, $\dot{K}_t$. Mathematically, $\dot{K}_t$ can be either positive or negative, but to be negative, we need a second-hand capital market in which we can unload usable equipment that we don't want. It is common in analyses to restrict $\dot{K}_t$ to be equal to or greater than zero, on the grounds that few second-hand capital equipment markets exist, leaving the only means of getting a smaller capital stock when you want one as setting gross investment, I_t , equal to zero and letting depreciation take its toll on the existing stock. This gives the relationship $\dot{K}_t = -\delta K_t$, which says that the capital stock is shrinking at the rate of depreciation. If only replacement investment is being made, $\dot{K}_t = 0$ and $I_t = \delta K_t$.

Third, we can "open up" this expression for investment demand, so we can "see what's inside it," by working with the full solution for the optimal quantity of capital: $K_t = K(w_t, c_t, p_t)$, which says that the optimal capital stock at each time t is a function of the labor cost, capital cost, and output price expected to prevail in that time period. Using this information on the determinants of the optimal quantity of capital, we can note that $\dot{K}_t$ is equivalent to $(\Delta K/\Delta c) \cdot (\Delta c/\Delta t)$: the time-rate of change of the capital stock can be broken into the change in capital stock in reaction to a change in the user cost of capital, times the time-rate of change of the user cost of capital. Now, recall the expression for the user cost of capital, and we see that the interest rate is an exogenous determinant of investment, but only through the user cost of capital. A higher interest rate will depress the rate of investment. Allowing that all the input and output prices can change over time (that is,

their time paths are not constant), we can get as a full expression of the investment demand function $I_t = I(w_t, c_t, p_t, \dot{w}_t, \dot{c}_t, \dot{p}_t)$, which says that investment is a function of the level of each input and output price, and the time-rate of change in each of them. This makes for a considerably more complicated relationship than those determining the current inputs, K_t and L_t , and current output, Q_t . As for other exogenous influences on the investment rate, a higher output price generally can be expected to raise the rate of investment, while a higher price of capital (not user cost, c , but capital goods price, q) and a larger depreciation rate (δ) both depress it. An increase in labor cost (or in any other factor price) has two effects that work in opposite directions: the substitution effect raises the investment rate, but can be swamped by a scale effect if the input cost increase is specific to a single firm. The increase in labor cost raises the firm's marginal cost at a constant product price, which causes the firm to reduce its output. If the labor cost increase is industry-wide, the entire industry would reduce output, but that supply reduction would raise the product price; since both $\Delta K/\Delta p$ and $\Delta I/\Delta p$ are positive, the scale effect is weakened, although the net effect still is likely to be negative: $\Delta I/\Delta w < 0$.

A final topic we should discuss in the theory of investment is that of adjustment costs. This refers to the cost, beyond the price of new capital equipment, of actually getting new investment "up and running." The typical observation is that more rapid investment is more costly than a more moderate pace of investment. For example, the cost of investing an amount of new capital equal to, say, $2I$ in time t would be more than twice as large as investing the amount I in each of two periods, t and $t + 1$. At the level of the individual firm, such increasing adjustment costs (increasing in the current investment rate) must be attributable to such factors as increased reorganizational costs, experienced primarily by labor, and increased "down time" for installation. At the level of an entire economy, more rapid investment could actually raise the base price of the capital goods to be installed as the economy works out a rising supply curve of capital goods.

In practice, the responses of investment to its determinants are sometimes delayed and sluggish and at other times quite rapid, even anticipatory as we noted above. The division of responsibility for these responses between rather vaguely appealed-to "adjustment costs" and "expectational effect" is not entirely satisfactory, but they are the best explanations we have so far. The expectational account is not quite as cheap as it may sound. When a firm observes, say, a price change for some critical input or one of its outputs, it may not know how much of that change is a relative change in that particular price, as opposed to a price-level change (a monetary phenomenon, in which all prices rise more or less together). It is also unlikely to know for certain whether price changes are likely to be permanent or transitory; similarly with many government policies such as taxation, which may actually contain a larger component of whim than prices determined in the economy.

8.5.2 Maintenance

Balancing the tradeoff between expenditures on equipment maintenance and the value of deferred productivity decay yields the concept of optimal maintenance. The particular actions taken with any given piece of equipment – do you oil it or keep it dry; do you take it out and air it or do you keep it covered? – involve engineering knowledge that is outside our current scope of calculations, other than how much they cost and how effective they are. A maintenance policy is a producer's plan, implicit or explicit, for taking actions to keep the productivity of her equipment and buildings at some desired level, which could be constant or declining over time.¹⁴ Viewed as a policy, a maintenance plan will be coordinated with the firm's production plan, which itself, of course, must be coordinated with its production plans.

The value of maintenance on any unit of equipment may depend, among other factors, on the age of the unit. The temporal pattern of productivity decay interacts with the remaining expected equipment life to produce a pattern of productivity improvement that a unit of maintenance expenditure (measured in money or consumable

commodities or the output of the equipment) generates. For instance, some equipment may experience little productivity decline in early years of use, while other types of capital goods may experience a sharp decrease initially, then a leveling out to a more gradual deterioration. Maintenance may be occasional rather than continuous; for example, some types of equipment may be overhauled every so many years, with little or no other attention in between overhauls. Meantime, the productivity of the equipment will decline between overhauls. However, the value of a current maintenance action will depend on the cumulative value of previous maintenance expenditures and their time pattern: the productivity increase derived from a current maintenance action will differ between a machine that has been neglected over its operating life and an initially comparable one that has been regularly maintained; and the productivity effect of such a maintenance action will differ between two identically treated pieces of equipment, according to when the last maintenance action was conducted. A machine that gets overhauled at regular, five-year intervals will be more productive in the second year after such maintenance than after four years; if the five-year cycle is really "optimal," the expense of the overhaul after two years would not be repaid by sufficient improvement in productivity to make the expense worthwhile.

The goal of a maintenance policy is to understand the response of equipment productivity to maintenance actions and plan the maintenance expenditures so as to get the largest value of increased productivity (prolonged equipment life) out of them. The prices of the goods produced with the equipment are important in determining the value of any equipment productivity improvement, so the same output price forecasts that were critical for investment decisions are important for maintenance decisions. An increase in output prices will raise optimal maintenance expenditures. Obsolescence of equipment – caused by technical progress in newer equipment – will reduce the value of the older equipment, lower optimal maintenance expenditure, and move the planned scrapping date of the older units closer to the present – and, of course, tip some units of old equipment over the edge into current scrapping.

Over cycles of varying product demand, maintenance will substitute for investment. A demand slump will cause gross investment (I_t in the notation of the previous subsection) to go to zero for some period of time, but $\dot{K}_t$ can be kept positive, or at least its rate of decrease can be reined in somewhat, by increased maintenance expenditure that operates on the depreciation term δ . Although we did not formulate δ as a function of anything in the last subsection, we can think of it as being some function like $\delta(m, M, a)$, where m is current maintenance expenditure, M is the cumulative previous maintenance expenditure, and a is an indicator of the age structure of the capital stock (or if we want to refer to a single piece of equipment, its age).

We can divide maintenance into two major types that depend on the type of problems to which particular types of equipment may be susceptible – preventive maintenance and preparedness maintenance. These two types of maintenance correspond to the sources of uncertainty in stochastically failing equipment. By stochastically failing equipment, we refer to equipment that maybe will or maybe won't work in each period of operation, only the operator may not be able to tell for certain without taking certain types of action. For instance, many types of capital have expected physical lifetimes – the one horse shay depreciation model highlights this type of failure – but the expected value (think of the statistical concept of expected value) has a distribution around it, so some units may fail much sooner than the operator expects and others may last considerably longer. Another type of uncertainty surrounding equipment failure is that it may be impossible to tell whether a particular unit will work the next time it is needed without taking some action – in other words, you can't tell just from looking at it whether it's still in satisfactory working order.

Preventive maintenance describes the types of actions taken on equipment whose condition can always be determined with certainty, because continuous operation of the equipment provides assurance of its operational state but its likelihood of failure during the next operating session is only known with some degree of uncertainty. The character of the failure likelihood is important, particularly whether it is constant or increasing over time. If the likelihood of failure is constant,

there is nothing to be gained by replacing a part before failure unless there are increasing costs of postfailure, relative to prefailure, replacement. An important question for maintenance policy¹⁵ is whether there are high costs of failure – that is, whether it's more costly to replace a unit of equipment after it's failed or sometime before, even if the unit still might be operational for a while yet. The replacement costs include the loss of operating time while the equipment is down, plus possible loss of material being processed by the equipment, and labor time. Possible maintenance actions under a preventive maintenance policy are planned replacement dates; planned inspection intervals, with repair or replacement if inspection reveals a deficiency; and monitoring, or continuous inspection.

Preparedness maintenance describes conditions where stochastically failing equipment is placed in storage and is called on for services only if a specific but predictable event (possibly characterizable as an emergency, but not necessarily) occurs. Several maintenance actions could be undertaken while such equipment is in storage. Inspections can be undertaken, but they may be imperfect in the sense that they may fail to recognize an existing defect or may indicate a defect or failure when the equipment is in fact in satisfactory condition. Inspections may be either periodic or sequential, the latter having the subsequent inspection scheduled after a given maintenance action has occurred. If the probability distribution of failure is increasing over time, each inspection serves effectively as a renewal of the equipment, equivalent to replacement. For equipment without such an increasing failure likelihood, a predetermined sequence of inspection intervals (the periodic inspection program) is superior to a sequential plan. As a radical example of costly postfailure repair or replacement amenable to the preparedness maintenance model, think of an ageing chariot: should we send it into battle even though some of its moving parts are getting old and might break under the strain of sudden, violent movement, because it would be expensive to replace it – or do we consider the cost of failure on the battlefield to be possibly unthinkable high, amply repaying the cost of current replacement?

Why go into such detail about maintenance actions with questionable applicability to ancient

capital equipment? Naturally, we believe there are many corresponding situations. Surely, the risk of fire with large inventories of either oils or grains was known to second-millennium bureaucrats – and probably even ordinary householders. Oil stores in Minoan and Mycenaean palaces surely were at risk of fire, from an earthquake spilling an oil lamp to the unfriendly visitor using the oil stores as lighter fluid, which appears to have happened at the House of the Oil Merchant at Mycenae (Wace 1951, 256; 1953, 13). At any rate, once the Minoan and Mycenaean palace stores caught fire, they generally burned down the entire edifices: for example, at Knossos (Evans 1921, 425, 453, 457–458, 462; 1935, 632–633, 648, 943) and at Pylos (Blegen and Rawson 1966, 10, 34). What actions might they have taken routinely to lower those probabilities? Are we willing to hypothesize, in our contemporary reconstructions of their lives and activities, that they consciously took no actions at all? Can we find any evidence regarding precautions, perhaps in inferences from Linear B archive tablets? Can any Mesopotamian archives shed light on maintenance activities of flammable inventories? The Kyrenia ship, which sank in the fourth century B.C. after an estimated 80-year service life, appears to have been the object of numerous maintenance and repair activities and, in the case of its keel cracking, subject to a fix-on-failure maintenance policy (Steffy 1985, 95–99; 1999, 396). Who would have put his products aboard such a tub?

8.5.3 *Scrapping and replacement*

Scrapping is the complete withdrawal of a piece of equipment from a firm's capital stock. A piece of equipment need not be physically broken for a producer to decide to scrap it. Equipment will be scrapped when its quasi-rent no longer covers its maintenance expense. An exception to this rule is the case of a temporary downturn in demand inducing a firm to withdraw some older equipment from current production but expectations of a reasonably near-term return to good business leads the firm to just idle the equipment rather than totally abandon it.

An alternative to scrapping unwanted equipment is to sell it in a second-hand capital market, but as we have noted on several occasions,

such markets in contemporary, industrialized economies are rare. Only vehicles and buildings generally find such markets. However, producers in contemporary developing countries do purchase and use second-hand equipment, so it may be unreasonable to suppose that second-hand equipment found no buyers in the economies of antiquity.

Scrapping is not a spur-of-the-moment decision but one that is planned at the time of acquisition of the equipment, in the form of an anticipated date of replacement. This is the same indicator we used above to characterize the durability of a piece of equipment, and indeed durability and planned scrapping date are alternative views of a single phenomenon. To the extent that postacquisition events (price and interest rate changes, or technical change, for example) affect quasi-rents in ways unanticipated at the purchase date, scrapping as an action obtains some independence from durability as a good's characteristic. Higher interest rates contribute to shorter optimal equipment life, as do higher rates of technical progress. Higher capital good prices and lower maintenance costs make for longer optimal life.

Scrapping of equipment naturally brings up the issue of replacement, and indeed, scrapping is unlikely to proceed or occur independently of plans for replacement – except when a production unit goes out of business. Replacement investment, defined by the term δK_t in our investment demand expression, is the purchase of equipment to maintain production capacity that is lost through capital decay and scrapping. It is common for replacement investment to exceed expansion investment in magnitude, and in times dominated by extended periods of low rates of technological change and slow growth in population and per capita income, such as probably characterized much – but possibly not all – of antiquity in the Near East and Mediterranean, replacement could be expected to be the dominant component of gross investment.

8.6 Consumption and Saving

Our general equilibrium treatment of capital theory has shown the relationships between

consumption and saving at a fairly abstract level. Just as section 8.5 described the richer context of investment when we had the luxury of describing producers in detail, this section will discuss a finer level of detail on the determinants of saving by consumers. We begin with a brief but alternative recapitulation of the intertemporal utility maximization problem that we introduced in Chapter 3, with a presentation that emphasizes the separate influences on consumption and saving. The following subsection follows up the important relationship between wealth and consumption highlighted by the intertemporal utility maximization exposition with two closely related hypotheses about the determinants of consumption. The final subsection discusses different ways that the individual consumption-saving relationship can add up to total, economy-wide savings-income ratios.

8.6.1 Intertemporal utility maximization

We will be quite brief with this exposition since we have already presented the analysis of intertemporal consumption decisions in Chapter 3. Our consumer has a utility function composed of his consumption in multiple periods: $u = u(c_0, c_1, \dots, c_n)$. His consumption over these periods will be restricted to the income he earns over the same periods (we abstract from “unearned” income, with little loss in generality but considerable loss in complication): $p_0 c_0 + a_1 p_1 c_1 + \dots + a_n p_n c_n = I_0 + a_1 I_1 + \dots + a_n I_n = W_0$; the p_t refer to the prices of consumption in the different time periods; the a_t represent a “flexible exchange rate” between consumption in period zero and each of the following periods, with $a_t = 1/(1 + i_{t-1})^t$, the familiar period-specific interest rate; W_0 is the present value of the future income stream, or the consumer's wealth evaluated in period zero. Now, we have the consumer maximize his utility subject to this present-value budget constraint: $\mathcal{L} = u(c_0, \dots, c_n) - \lambda(\sum_t a_t p_t c_t - W_0)$. Next we vary the size of c_t to obtain the first-order conditions for maximization, which are $(\Delta u / \Delta c_t) / a_t p_t = \lambda$, or in words, the ratio of the marginal utility of consumption in time period t to the discount-weighted price of consumption in that period is equal to the marginal utility of

wealth, λ . Since we have a similar first-order condition for consumption in each time period, we have the ratio of marginal utilities in each time period equal to the ratio of the corresponding, discount-weighted prices of consumption: $MU_i/MU_j = a_i p_i / a_j p_j$. A wealth-compensated reduction in $a_i p_i$ would increase consumption in period i , c_i , and decrease consumption in the other periods j as a group, $\Sigma_j c_j$; and an increase in wealth, W_0 , would increase consumption in all the periods since the weighted average of the individual-period wealth elasticities equals one. From this set of relationships, we get a demand function for consumption in each period; letting $p_i = p_j$ for every period, we get $c_t = f(1/a_{t+1}, W)$. This demand function for composite consumption in a particular time period, in terms of the interest rate and wealth rather than in terms of either prices and income or prices and utility, is commonly called a consumption function. An increase in $1/a_t$ (equivalent to a rise in the interest rate) reduces consumption (that is, $\Delta c_t / \Delta(1/a_t) < 0$); and an increase in wealth raises consumption by an amount less than the increase in wealth ($0 < \Delta c_t / \Delta W < 1$; because the increase in wealth increases consumption in all periods, the increase in any particular period's consumption has to be less than the increase in wealth).

Now, for each period, we have the simple definition that savings are just the difference between income and consumption: $s_t = I_t - c_t$. Writing savings as a function, we have $s_t = g(1/a_{t+1}, W, I_t)$. Compare the arguments in the consumption function and the savings function and notice that they are different: consumption is a function of wealth, which changes only slowly, while savings depend on wealth *and* current-period income. This set of relationships forms the basis of two important hypotheses about consumption both at particular points in time and over time: the lifecycle saving hypothesis and the permanent income hypothesis. Although one of the models refers to saving and the other names income as its subject, they both make predictions about the behavior of consumption and saving as functions of wealth, particular measures of income, and stage in the individual or family life cycle. We turn now to those two hypotheses.

8.6.2 Hypotheses about consumption

The permanent income hypothesis is a model of the relationship between income and consumption – total consumption in a time period, not consumption of some particular good. The model begins with the observation that income fluctuates from year to year, but that consumption appears to vary somewhat less. Observed income, y , is composed of two parts, permanent income, y^p , and transitory income, y^T . Permanent income is a wealth measure more than a current income measure; it is roughly the income that an individual expects to be his or her “normal” level of income, abstracting from year-to-year, transient fluctuations, for some number of years into the future. Transitory income can be either positive or negative. Paralleling permanent and transitory income is the division of observed consumption into permanent and transitory components, denoted by c^p and c^T . Permanent consumption is proportional to permanent income, the factor of proportionality depending on the interest rate, the consumer's age, family structure, and possibly other variables reflecting the ability to command resources. Transitory consumption is not systematically related to either permanent or transitory income.

Then, in any period, observed consumption will be a proportional factor of permanent income, plus transitory consumption: $c = k(i, a, f, u) \cdot y^p + c^T$; in the k function, i represents the interest rate, a represents the consumer's age, f a measure of family structure (marital status, number and age of children, other individuals in the household, and so forth), and u indicates that some other variables might be included. The marginal propensity to consume out of permanent income (the incremental change in consumption expenditure per increment in permanent income) will be equal to the factor k , while the marginal propensity to consume out of transitory income is zero (which does not mean that no transitory income is ever spent, only that on average, over a number of years of receiving positive and negative transitory income, the expected expenditure out of a unit change in it is zero).

Thinking back to the definition of saving, and the corresponding consumption and saving

functions we derived in the previous subsection, we can see where the lion's share of transitory income goes: into saving. However, we must be careful how we define consumption. The purchase of a durable good out of positive transitory income would be classified as an act of saving rather than as a consumption good purchase, because the durable good is a capital unit that provides consumption services over a period of time.

An individual's assessment of his or her permanent income at any particular time is subject to revision in light of changes in either current income or in long-term prospects. However, the effect of a single year's income on the assessment of permanent income is likely to be limited because initially the recipient will be unsure what part of it is permanent and what part is transitory. Consequently, the assessment of permanent income is revised slowly, and the adjustments to consumption are made at a corresponding pace. How would an individual adjust her permanent income and her consumption to the receipt of an inheritance? Assuming that she always knew that she would receive the inheritance, she would have incorporated its expected value into her assessment of permanent income already – from the moment she understood she would receive it. Its full effect on consumption would have been a longstanding component of her consumption. Any difference between the expected value of the inheritance and its actual value she would assign to transitory income, either positive or negative. The date she actually receives the inheritance makes no difference to either her permanent income or her permanent consumption.

The permanent income model also predicts that consumption will fluctuate less, over time, than does income or output, which accords well with empirical observations. However, limitations on borrowing, because of weakly developed financial institutions or impediments such as lack of collateral, can limit the predictive success of the model without further adaptation to institutional conditions. It may appear superficially that the permanent income model of consumption gives no scope for risk aversion as a rationale for saving (that is, not consuming),

but recall the zero marginal propensity to save out of transitory income. Knowing implicitly that he will experience about as many fluctuations below the average (permanent income) as above, the consumer will tend to save most of positive transitory income, if for no other reason than to have it available to consume when transitory income is negative.

The lifecycle savings model also takes a long view of the income / wealth determinant of consumption, but focuses somewhat more directly on the saving behavior. Using an annuity-like version of income, based on expected lifetime earnings, it predicts negative saving during early years, positive saving during the mature, productive years, and dissaving again during the retirement or declining years. Figures 8.18(a) and 8.18(b) show the time pattern of income, consumption, and saving predicted by the model. The critical determinant of consumption in the lifecycle model is the ratio of current income to wealth, with the present value of expected future income included in wealth; the telling measure of consumption is the ratio of current consumption to current income, c_t/y_t . If we considered two individuals with identical current incomes and identical wealth but the income of one of them was subject to greater year-to-year fluctuation, that individual would consume less than the one with the more stable income. Although there is a positive correlation between age and income in general, looking across different income groups and finding consumers of the same age but with different incomes, a younger person would have a higher c_t/y_t than a middle-aged person, and probably about the same as an older, retired person. Larger family size raises the c_t/y_t ratio, but the relationship is small and nonlinear.¹⁶

Again, the institutional mechanisms that facilitate the borrowing and lending pattern will vary across societies, ranging from intrafamily transactions to village money lenders to true banking systems. Additionally, government programs in many contemporary societies offer some support in the retirement years, as well as various forms in the earlier years (through, for example, public schooling); public (government, state, or whatever term is desired) support in individuals'

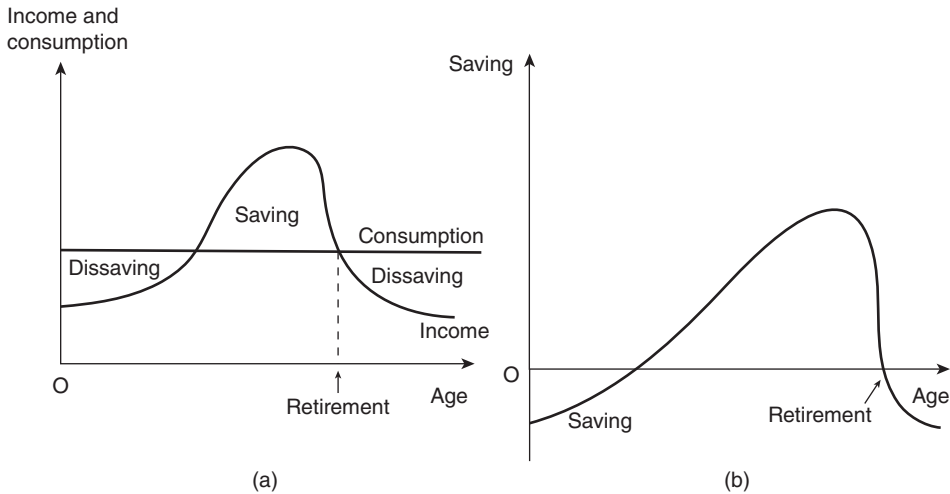


Figure 8.18 (a) Income and consumption over the lifecycle. (b) Saving over the lifecycle.

declining years, at least in towns and cities, may be worth looking for in the textual records of the ancient Near Eastern societies and Egypt (the state of records in the Bronze Age Aegean may preclude any fruitful search there).

While we may be accustomed to thinking of saving in terms of money squirreled away at a bank, we should be more open minded about the forms saving is likely to have taken in the ancient Mediterranean lands. We have already noted durable goods as a form of saving. The liquidity (the ability to sell an asset quickly without having to accept a drastic discount on its value) of particular durable goods must have been a concern. For instance, adding on a room to a house might not have been particularly liquid investment but the ability to rent the room still would have yielded income. Saving up bricks might sound odd, but if bricks were in relatively constant demand, and users weren't picky about sources, they could have been a viable form of savings. In shanty towns of some of the contemporary world's developing countries, it is common to see wooden houses with uneven lengths of board forming the outer walls: the extra lengths of board are forms of saving – as long as you don't cut off the extra lengths of board, you'll know where they are when you want to use them for something else (Perlman 1976, 38–39, plates 6–7)!

The observation that expenditures on children may contain a sizeable element of saving leads to consideration of the role of intrafamily transfers within multigenerational families such as characterized much of the ancient world. With limited external financial institutional structures to facilitate flows of consumption between lenders (savers) and borrowers (either investors or consumers), such intrafamily, intergenerational transfers would have served as the primary means of consumption of the elderly, even considering the relatively youthful age structure of ancient populations. Most of these transfers would have been made within the same period the consumption goods were produced rather than in the form of stores of value of varying degrees of liquidity. This type of saving system has the structure of what has been formulated as the "overlapping generations model." In one version of that model, each generation lives for two periods, the first being a period of production, the second being a period of retirement, but consumption must occur in both periods. In the first period of life, individuals combine their labor with the capital of the people who are currently in their second period; they save part of their labor income and carry it forward to their second period of life. Individuals in the second period of life consume both their capital income and

their existing wealth (that is, they “eat” their capital). A rise in the interest rate need not increase the saving of the young – the income effect of having more second-period consumption can offset the substitution effect of the improved relative price of second-period consumption. If relative risk aversion is high (greater than one), the income effect dominates; if it is less than one, the substitution effect dominates.¹⁷

8.6.3 *Individual and aggregate savings*

Saving at the level of the individual consumer may differ from saving at the level of the entire economy. For example, the lifecycle model predicts that with a stationary population and zero growth of income per capita, the saving of younger birth cohorts will be exactly offset by the dissaving of the older birth cohorts, and the aggregate savings rate will be zero. If life cycle effects dominate saving behavior, the aggregate savings rate will be heavily influenced by the population growth rate. A spurt in population growth would yield a relatively large age cohort with a low or negative savings rate; providing capital for them to work with in their mature years would require an increase – possibly a large increase – in the savings rate of the cohort preceding them; otherwise they will spend their working years with a lower capital-labor ratio than the preceding cohorts and consequently experience lower per capita incomes. If this new, larger cohort saved at the same rate as the preceding cohorts’ rates, the capital so formed would not appear in time to raise their own, working capital-labor ratios to previous levels, and they in fact would save a smaller amount per head than their predecessors.

Different problems occur with falling population growth rates (not a negative growth rate, but a reduction in the rate at which the population increases). The cohort whose birth rate (or survival rate, or whatever) dropped will be followed by a relatively smaller cohort. If retirement consumption is financed in the current period by the working cohort, the following cohort’s saving rate will be depressed, yielding a lower capital-labor ratio for the cohort following them. If the older cohort (the one whose birth rate dropped) accumulated capital during their working lives in the form of claims on future consumption, they

will drive up the price of consumption goods produced by the smaller, following cohort.

High population growth rates in general require working-age members of the population to save higher proportions of their incomes in order to maintain the capital-labor ratio constant for the succeeding working generation, although there is no implication they consciously would do that. In general, the savings rate would not rise sufficiently to accomplish this, and the succeeding generation would work with a lower capital-labor ratio and produce lower per capita incomes.

8.7 Capital Formation

Investment results in the formation of new capital, but not all new capital expands the size of the capital stock.¹⁸ Capital can be divided into two major categories, depreciable and nondepreciable. While depreciation, or capital consumption, is a common contemporary accounting term with little economic content, economic depreciation is the decrease in value of capital due to wear and tear and technological obsolescence, allowing for normal maintenance expenditures, which may increase as capital ages. Nondepreciable capital never loses its value and has an indefinitely useful economic life. Land is the principal category of nondepreciable capital, although land need not be nondepreciable because its productivity can be degraded, sometimes permanently, by types of cultivation, and its physical mass can be reduced through erosion. Land is a small component of contemporary nations’ capital stocks, although it may have represented a larger proportion of the value of societies’ capital stock in various places and times in antiquity.

Focusing on depreciable capital, depreciation of existing stock must be subtracted from gross additions to the stock in a period to arrive at the net addition to the capital stock, or net capital formation. The current value of a society’s depreciable capital stock is the sum of the gross increments at many previous dates less the depreciation of those gross stocks at all dates between the formation of the oldest components of the stock and the present. With some simplifications, some general formulations can be derived. Call the total output of goods and services in an economy its gross national product (GNP). Suppose

that gross capital formation is a constant fraction f of GNP each year. Second, suppose that GNP has been growing at a constant, nonzero rate of g each year. This growth rate g could simply be the growth rate of the population, yielding a zero per capita growth rate but at any rate, this assumption will be relaxed below. Finally, suppose that the technology of capital is unchanging and that units of capital last n years, with straight-line depreciation between years 0 and n , such that capital today that is n years old has a stock equal to its initial magnitude times $1/n$; capital that is one year younger has a current stock equal to its initial magnitude times $2/n$; and so on, until we get to capital that was installed last year, which has a current stock equal to n/n of its initial magnitude.

Using these three simplifying assumptions, the current value of the capital that was formed n years ago is $fG_0/(1+g)^n$, where G_0 is the current period's GNP, and the remaining proportion of its original value is $1/n$. We don't need to go to capital formed $n+1$ years ago because none of it is left today. In the first expression, dividing the fraction of current GNP that is devoted to gross capital formation by $(1+g)^n$ shrinks today's GDP down to what it would have been n years ago at the constant growth rate g , yielding the gross capital formation at that date. Multiplying the first expression by $1/n$ reduces the original stock by the amount it has depreciated over n years with straight-line depreciation. Moving forward in time, the gross formation of capital one year closer to the present is $fG_0/(1+g)^{n-1}$, and $2/n$ of its original stock remains today. Forming the expressions for gross investment for each of the years between n years ago and 1 year ago, and summing them yields an expression for the value of current capital stock: $C_0 = \frac{fG_0}{g} \left[\frac{ng-1+(1+g)^{-n}}{ng} \right]$.¹⁹ Depreciation, or capital consumption, at the current date is $D_0 = \frac{fG_0}{ng} [1 - (1+g)^{-n}]$. The share of depreciation in gross capital formation then is $D_0/fG_0 = [1 - (1+g)^{-n}]/ng$.

When the rate of economic growth is slower, or when the economic life of depreciable capital is shorter, the share of depreciation in gross capital formation is larger; both conditions together strengthen this effect. Stated alternatively, the ratio of net capital formation to gross capital

formation is smaller when the growth rate is smaller or capital life is shorter. Although maintenance has not been included in the formulation here, higher maintenance costs, which remain outside this accounting model of capital formation, would further decrease the net productivity effects of having more capital. The intuition of a shorter life span of capital leading to a higher share of depreciation in gross capital formation is straightforward: with a shorter lifespan, each year a larger proportion of the original value of capital is lost to depreciation. The intuition behind the growth rate's effect is somewhat more subtle but still quite accessible. With a faster growth rate, previous years' GNP would be smaller, and the gross capital formation associated with a fixed fraction of it would be smaller, so the fraction that is lost to depreciation in the current period is smaller.

These conditions almost certainly prevailed across wide periods and regions in antiquity. Growth rates, both per capita and aggregate – that is, allowing for population growth – are believed to have been very low or static over lengthy periods, excepting periods of recovery from generalized societal disasters such as the end of the Mycenaean civilization and a possible several-century period of prosperity during the early Roman Imperial Period. Considering the materials from which capital implements – machinery – and many structures – private buildings in particular; public utilities such as dams, aqueducts and roads seem to have been more durable – were made, their economic life spans probably were fairly short, especially compared to capital developed in the past several centuries. Nonetheless, research on the effective life span of various types of capital equipment could be illuminating. It would not be difficult to suspect that maintenance expenses of implements and private structures represented a larger proportion of their gross output in antiquity than has been the case in recent centuries, but I am not aware of evidence on the subject. The upshot of any of the three of these conditions, one of which (low growth rates at the best of times) was certainly the case, is that growing the capital stock would have been difficult in antiquity. A large proportion of gross investment simply would have replaced capital consumption from the previous period. It is easy to see that

during periods of negative growth, a special case of the nonzero g expressions above, capital consumption would have exacerbated problems societies faced.

When $g = 0$, the expression for the cumulative proportion of previous years' capital stocks that remain today is $(n + 1)/2$. The expression for gross capital formation n years ago when $g = 0$ is simply fG_0 , since GNP n years ago is the same as it is currently. Current capital stock, debiting depreciation, is $nfG_0(n + 1)/2$.

8.8 Suggestions for Using the Material of this Chapter

The models of this chapter, to a considerable extent, should affect a scholar's view of the creation and use of capital equipment and structures. Capital observed in one period is related to consumption and saving in previous periods, the technology for producing the capital, and the productivity of the capital once produced. The existence of large quantities of capital in a society indicates that its people have been producing more than they've consumed for some extended period of time. And since capital items are expected to last for some time, they are votes for the future, an insight into the optimism or pessimism of people in antiquity.

Very scarce capital implements imply high interest, or discount, rates: the present is far more important than the future in people's planning. For archaeologists, comparing the scarcity of capital implements between sites will be difficult, with controls difficult to apply. Nonetheless, keeping in mind that capital-poor sites suggest high discounting of the future may help in interpreting other facets of a site that I simply do not anticipate.

Do the Egyptian royal tombs imply low discount rates? A relatively high valuation of the future relative to the present? Or do they represent, at least in part, the low productivity of savings – capital – invested in alternative types of capital? To what extent can changes in funerary architecture, Egyptian or Minoan or Mycenaean or Assyrian or Roman, be associated with changes in the productivity of capital invested in other forms versus changes in income or completely unrelated changes in society?

Monumental architecture, such as the Minoan and Mycenaean palaces, the Egyptian temples, pyramids and monuments of specific pharaohs, the Near Eastern ziggurats and other religious structures, represent major capital items of those ancient societies. We know that the Minoan and Mycenaean palaces were constructed over distinct periods, with additions, occasional clearances, and rebuilding after various destructions. Much, if not most, of the cost of these capital items would have been in the form of labor, assuming that the stone was quarried without royalty payments and that the timbers may also have been free for the taking – and possibly with overcutting, treated in Chapter 13. Theoretically, we know that capital comes from saving, which is, roughly speaking, current production minus current consumption. The saving would have been in the form, possibly, of farm labor beyond that actually employed in farming activities, possibly supplemented by nonfarm labor with different skills, diverted from its usual employment. It is tempting to assume a simplified story to the effect that whoever was in charge – king, pharaoh, wanax, or whatever he or she was called – just told these workers to show up and build, and they did; effectively, he or she “owned” everything and everybody and just ordered them around. In this simplified explanation of where the resources came from, these political leaders did not borrow from a capital market – float a government bond – to finance their construction projects because there wasn't one. So does this mean that the fellow in charge (FIC) had to create this monumental architectural capital out of current savings only rather than accumulated previous savings, since labor doesn't “keep” as would, say, squirreled-away gold, which could then be traded for labor services or imported construction items? An alternative story would suggest that the FIC didn't actually own everything and everybody,²⁰ but that there were competing or countervailing power centers in the form of at least notionally subject aristocrats (not a foreign concept in the literature on these times and places), and that the FIC actually had to pay the commoners to get any work out of them – the lash alone simply was too inefficient. Now, in this setting, the FIC talks with his nominally subordinate aristocrats and persuades them to pay their own subjects to work on his monumental architectural project,

in return for which, the FIC agrees to give these aristocrats something that they want for the next so many years.²¹ The aristocrats may be able to divert the time of some of their subjects from working on some of their activities (possibly agricultural work, possibly his or her own construction projects) and as a consequence have a smaller value of current production with which to buy (“exchange,” “trade,” “gift-exchange,” or whatever) attractive imports from overseas. The FIC’s offer to give them something over a period of time to make up for this lost production is sufficient to get them to tighten their current-consumption belts for a while. According to this alternative explanation of the financing of monumental architecture, although the resources used in constructing the new capital come out of current production rather than a stock of accumulated previous production, the FIC is still borrowing to build. It would be easy to imagine the FIC tapping into some form of stored wealth the aristocrats own to purchase labor services or imported materials that go into the construction. In either case, the FIC has done the equivalent of floating a government bond, the bondholders

including the aristocrats in the polity and possibly some of their subjects as well as some subjects of the FIC himself. Interest rates appear, and some surrogate for a capital market emerges, even if it appears a far cry from an organized exchange. This alternative story could serve as an hypothesis, or a set of hypotheses, with which to confront various texts from Egypt and the Near Eastern kingdoms. The Minoan period is absent such texts, and the Linear B texts of the Mycenaean realms are probably too limited to shed much light on construction, and indeed typically come from times when destruction was more the rule than construction. Some information regarding pace of construction might be teased out of material remains that could shed light on the proportion of any period’s total production that could have been devoted to monumental construction; at any rate, I’ll not underestimate the ingenuity of the archaeologists in finding surrogate indicators of concepts in which they’re interested.

I hope these suggestions prompt ideas well beyond my own inklings of applications.

References

- Aldrete, Gregory S., Scott Bartell, and Alicia Aldrete. 2013. *Reconstructing Ancient Linen Body Armor; Unraveling the Linothorax Mystery*. Baltimore MD: Johns Hopkins University Press.
- Berlin, Ira. 1998. *Many Thousands Gone: The First Two Centuries of Slavery in North America*. Cambridge MA: Harvard University Press.
- Blegen, Carl W., and Marion Rawson. 1966. *The Palace of Nestor at Pylos in Western Messenia*. Volume I. *The Buildings and their Contents*. Princeton NJ: Princeton University Press.
- Bliss, C.J. 1975. *Capital Theory and the Distribution of Income*. Amsterdam: North-Holland.
- Brady, Dorothy. 1956. “Family Saving, 1888 to 1950.” In *A Study of Saving in the United States*, Vol. III. *Special Studies*, edited by Raymond Goldsmith, Dorothy Brady, and Horst Mendershausen. Princeton NJ: Princeton University Press, pp. 137–273.
- Brewer, Douglas J., and Emily Teeter. 1999. *Egypt and the Egyptians*. Cambridge: Cambridge University Press.
- Buchholz, Hans-Günter, and Joseph Wiesner. 1977. *Kriegswesen, Teil 1. Schutzwaffen und Wehrbauten*. *Archaeologia Homerica* E1. Göttingen: Vandenhoeck and Ruprecht.
- Crouwel, J.H. 1981. *Chariots and Other Means of Land Transport in Bronze Age Greece*. Amsterdam: Allard Pierson Museum.
- Crouwel, J.H. 1992. *Chariots and Other Wheeled Vehicles in Iron Age Greece*. Amsterdam: Allard Pierson Museum.
- Evans, Arthur. 1921. *The Palace of Minos. A Comparative Account of the Successive Stages of Cretan Civilization as Illustrated by the Discoveries at Knossos*. Volume I: *The Neolithic and Early and Middle Minoan Ages*. London: Macmillan.
- Evans, Arthur. 1935. *The Palace of Minos. A Comparative Account of the Successive Stages of Cretan Civilization as Illustrated by the Discoveries at Knossos*. Volume IV, Part II: *Camp-stool Fresco, Long-robed Priests and Beneficent Genii; Chryselephantine Boy-God and Ritual Hair-Offering; Intaglio Types, M.M. III – L.M. II, Late Hoards of Sealings, Deposits of Inscribed Tablets and the Palace Stores; Linear Script B and its Mainland Extension, Closing Palatial Phase; Room of Throne and Final Catastrophe*. London: Macmillan.
- Forsdyke, Sara. 2012. *Slaves Tell Tales; And Other Episodes in the Politics of Popular Culture in Ancient Greece*. Princeton NJ: Princeton University Press.

- Friedman, Milton. 1957. *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.
- Isager, Signe, and Jens Erik Skydsgaard. 1992. *Ancient Greek Agriculture; An Introduction*. London: Routledge.
- Knight, Frank H. 1944. "Diminishing Returns from Investment." *Journal of Political Economy* 52: 26–47.
- Kuznets, Simon. 1955. "International Differences in Capital Formation and Financing." In *Capital Formation and Economic Growth*, edited by Moses Abramovitz. Princeton NJ: Princeton University Press, pp. 19–106.
- Kuznets, Simon. 1973. "Capital Formation in Modern Economic Growth (and some implications for the past)." In Simon Kuznets, *Population, Capital, and Growth; Selected Essays*. New York: Norton, pp. 121–164.
- Marinatos, Nanno. 1993. *Minoan Religion; Ritual, Image, and Symbol*. Columbia SC: University of South Carolina Press.
- McKeown, Niall. 2011. "Resistance among Chattel Greek Slaves in the Classical Greek World." In *The Cambridge World History of Slavery*. Volume I. *The Ancient Mediterranean World*, edited by Keith Bradley and Paul Cartledge. Cambridge: Cambridge University Press, pp. 153–175.
- Morgan, Edmund S. 1998. "The Big American Crime." *The New York Review of Books* 45 (19): 14–18.
- Morgan, Philip D. 1998. *Slave Counterpoint: Black Culture in the Eighteenth-Century Chesapeake and Lowcountry*. Chapel Hill NC: University of North Carolina Press.
- Olesen, John Peter, ed. 2008. *The Oxford Handbook of Engineering and Technology in the Classical World*. Oxford: Oxford University Press.
- Pertman, Janice E. 1976. *The Myth of Marginality; Urban Poverty and Politics in Rio de Janeiro*. Berkeley, CA: University of California Press.
- Petrie, W.M. Flinders. 1917. *Tools and Weapons, Illustrated by the Egyptian Collection in University College, London, and 2000 Outlines from Other Sources*. London: Constable & Co.
- Postgate, J.N. 1992. *Early Mesopotamia; Society and Economy at the Dawn of History*. London: Routledge.
- Soles, Jeffrey S. 1992. *The Prepalatial Cemeteries at Mochlos and Gournia and the House Tombs of Bronze Age Crete. Hesperia Supplements*, Vol. 24. Princeton NJ: American School of Classical Studies at Athens.
- Steffy, J. Richard. 1985. "The Kyrenia Ship: An Interim Report on Its Hull Construction." *American Journal of Archaeology* 89: 71–101.
- Steffy, J. Richard. 1994. *Wooden Ship Building and the Interpretation of Shipwrecks*. College Station TX: Texas A & M University Press.
- Steffy, J. Richard. 1999. "Ancient Ship Repair." In *Tropis V: 5th International Symposium on Ship Construction in Antiquity, Nauplia, 26, 27, 28 August 1993: Proceedings*, edited by Harry Tzalas. Athens: Hellenic Institute for the Preservation of Nautical Tradition, pp. 395–408.
- Toynbee, J.M.C. 1971. *Death and Burial in the Roman World*. Baltimore MD: Johns Hopkins University Press.
- Van de Mieroop, Marc. 1992. *Society and Enterprise in Old Babylonian Ur*. Berlin: Dietrich Reimer.
- Wace, A.J.B. 1951. "Mycenae, 1951." *Journal of Hellenic Studies* 71: 254–257.
- Wace, A.J.B. 1953. "Mycenae, 1939–52. Introduction." *Annual of the British School at Athens* 48: 3–18.
- Westermann, William L. 1955. *The Slave Systems of Greek and Roman Antiquity*. Philadelphia: The American Philosophical Society.

Suggested Readings

- Burmeister, Edwin. 1980. *Capital Theory and Dynamics*. Cambridge: Cambridge University Press. Chapter 2.
- Hirshleifer, J. 1970. *Investment, Interest and Capital*. Englewood Cliffs NJ: Prentice-Hall. Chapters 2–4, 6.

- Pindyck, Robert S., and Daniel L. Rubinfeld. 2001. *Microeconomics*, 5th edn. Upper Saddle River NJ: Macmillan. Chapter 15.

Notes

- 1 Of course, there will be many cases in which the mistreated equipment must be simply junked: for example, a stone grinding wheel that breaks, a building that burns down because of neglect.
- 2 Recent studies of American slavery in the eighteenth and nineteenth centuries C.E. have begun

to retrieve some of these slave-master negotiations from the shadows of marginal documentation, and even to estimate some of their major social and environmental (broadly speaking) determinants. See P. Morgan (1998), Berlin (1998), and the discussion in E. Morgan (1998). Most

- recently, and in the Greek context, Forsdyke (2012, Chapter 2, 173–174) has explored the role of fables and other components of popular culture in maintaining some social equilibrium among slaves and between slaves and masters.
- 3 See Marinatos (1993, Chapter 2) on an interpretation of Minoan customs involving the use of funerary monuments; Soles (1992, 236–242) reports shrines outside tombs and in nonburial rooms inside tombs on Bronze Age Crete; Brewer and Teeter (1999, Chapter 6) on Egyptian customs; Toynbee (1971, Chapters 2 and 3) on Roman customs.
 - 4 I say “largely” because, to the extent that loan contracts were “rediscounted” and passed around as tokens that third and fourth parties could use to acquire consumption goods, there were at least fiat components to ancient societies’ money stocks.
 - 5 I have in mind as possible monetary stocks of grain the stores of grain known to have been held in ancient Mesopotamian cities, as well as the possible grain stores that scholars long have associated in one fashion or another with the koulouras on the west courts of the Minoan palaces (and the south side, of course, at Mallia). Some portion of these stocks may have represented “normal” consumption inventories – that is, holdings for anticipated consumption during the coming year – another portion may have represented what would be called strategic stocks today – that is, emergency stocks – but altogether leaving some amount of the stock holdings to be attributed to monetary uses, possibly as regular wage and salary payments to personnel on the palatial “payrolls.”
 - 6 This abstracts from the important service of “liquidity” that a money stock provides, a dearth of which can have “real” effects in terms of the employment of available resources and the consequent production of consumption goods. We will discuss this issue in the chapter on money and banking.
 - 7 This sum is commonly called the present discounted value or just present value. It is not a net present value in this particular example simply because we haven’t admitted any costs of using the capital to produce the income (such as maintenance or depreciation) that must be debited from the gross income the apartment building yields.
 - 8 If we wanted to allow for different rentals in each year and different interest rates, we could represent the accumulation of those time-varying parameters with an integral expression, which we won’t do here.
 - 9 See, for example, Petrie, (1917, 57, pl. LXXIII, no. 77). We might as easily refer to one of the smith’s coal shovels from the thirteenth-century B.C.E., found on Cyprus and Sardinia.
 - 10 This is Frank Knight’s (1944) Crusonia Plant model, which takes an inventory view of capital: if left alone, the Crusonia Plant will continue to grow; if desired, the consumer can cut off a bit and consume it directly.
 - 11 This equation makes two assumptions: (i) that the investment in the form of deferred consumption creates an equal amount of new capital (that is, that there are not decreasing marginal returns in investment in the production of the capital good); and (ii) no depreciation of the capital good.
 - 12 This tradeoff being a variable in the model implies that time preference is endogenous in the same sense that the quantity demanded of a good may have a unique, equilibrium value at the same time the consumer has an entire schedule of demanded quantities, of which only one will be selected, given supply conditions.
 - 13 I won’t go into the exact character of these problems here, but the general “problem” is that the recommendations about the best set of decisions can be sensitive to timing characteristics of the project. If you’re not careful and use the net present value criterion mechanically, you might fail to identify the project (in terms of quantity and timing of investments) that yields the largest social (or private) benefit.
 - 14 We cannot rule out the occasional desire to actually raise the productivity of existing equipment through maintenance and repair, usually as an alternative to replacement or not producing at all.
 - 15 Or replacement policy, the subject of the next subsection; in the use of the term in this section, we have in mind more replacement of parts than entire pieces of equipment.
 - 16 Brady (1956, 211). See also Friedman (1957, 122–123), for discussion of Brady’s results. Brady found that the consumption-income ratio rose in proportion to the sixth root of the number of family members. Thus a family of five would consume 5% more of its income than a family of 3.5; if the 3.5-person family consumed 88% of its income, the five-person family would spend 93%. However, these consumption figures do not take account of the fact that a substantial amount of the expenditure involved in raising a child may be viewed as investment in human capital. This consideration would reduce the observed ratio of consumption to saving for both families but by a larger proportion for the larger family.
 - 17 Recall the coefficient of relative risk aversion from Chapter 3: from the utility function $u = u(c_t)$, relative risk aversion, which is a numerical

measure of the relative curvature of the utility function – that is, the curvature of utility, u , as consumption, c_t , increases – is $-cu''/u' > 0$. Our notation is as follows: u' denotes the marginal utility of consumption, $\Delta u/\Delta c_t$, and u'' denotes the change in marginal utility as c_t increases: $\Delta(\Delta u/\Delta c_t)/\Delta c_t$. If $u'' = 0$, the “curve” of utility as a function of consumption is a straight line, and we have risk neutrality. Any value greater than zero indicates some degree of (relative) risk aversion.

- 18 This section draws on Kuznets (1955, 76–81; 1973, 136–161).

- 19 This is called the perpetual inventory method of measuring current capital stock.
- 20 Or if he notionally did, he couldn't have made the claim on the ownership effective without angering the subjects and risking upsetting the power balance in the society; that is, he would have lost political legitimacy in attempting to enforce a notional right.
- 21 At a superficial reading, this might sound like a gift-exchange system, but I do not have in mind the agonistic character involved in gift exchange – a simple *quid pro quo* rather than an effort at besting one another.

Money and Banking

If we find evidence, or otherwise infer, that the quantity of coinage increased in a particular region over a certain period of time in antiquity, what could we say about the possible causes or consequences of that increase? Monetary theory offers a number of possibilities for exploring these issues.

Money is a commodity, or at least an item of choice, just like any other good in an economy, from apples to shirts to houses. There is a demand for and a supply of money, just as there is for every other good and service in an economy. Having said that, it is a peculiar commodity in that, at least in its use as money (many moneys can be used for other purposes at the same time that some units of it are used as money), it is not valued at all for itself, only for what it will let its holder acquire with it. A society's or economy's money supply is a stock – a quantity at a particular time – but flows of money – the number of units of it used during a period of time – are equally important. While the stock of, say, shovels in the economy at any given time will affect the real production possibilities of the economy – how much the people can produce – the size of the stock of money is virtually immaterial to the quantity of services it provides to the economy. Its mere existence is what matters. But, to demonstrate the scope of complexity, *changes* in the size of

the money stock can have immensely important consequences for the entire economy – although when all is said and done, all adjustments to the new money stock accomplished, it is not necessarily true that the configuration of real production, real allocations, and real relative prices will be much different from what they were with the previous money stock. And it can be created by banks as well as by monetary authorities.

Considering these paradoxical properties of money, it is not surprising that it has taken economic thinkers hundreds, possibly even several thousands, of years to reach the present state of limited understanding and agreement about the details of monetary theory. Nevertheless, there is a reasonable core of agreement in the theory, sometimes expressed in different forms by scholars adopting different perspectives. I endeavor in this chapter to present something close to this core of agreed-upon monetary economics, although my own intellectual leanings undoubtedly will assert themselves where some choices are necessary.

I keep firmly in mind that many of the monetary and transaction technologies and institutions under whose influence contemporary monetary thought has been developed did not exist in antiquity. For example, contemporary monetary

theory finds it important to include a bond market in analyses of the demand for money; the theory of money supply has been developed in the context of fractional reserve banking; and the most contentious elements of monetary theory over the past century have centered on the effects of monetary changes on industrialized employment and investment. I will be sparing in discussing these topics, but there is a risk in not discussing them at all: despite the vast differences in institutional trappings, many of the same activities as those accomplished through bond markets and the creation of money and investment lending by banks may have been conducted in different fashions in ancient societies. Some certainly were. Even where certain institutional features were absent, it will prove useful to have a basic understanding of what their absence caused, in the way of both compensations and preclusions.

9.1 The Services of Money

The subject of this section is what is traditionally called the “functions of money.” I use the term “services” rather than “functions” to emphasize that money actually produces something that is valued, several things, in fact. These three services are serving as a medium of exchange, as a store of value, and as a unit of account. I discuss each in turn.

9.1.1 *Money as a medium of exchange*

In a sufficiently compact setting – few people, short distances, small numbers of products – people who possess one thing but want another can find a person with what they want and, with sufficient divisibility in each good, can reach a ratio of goods acceptable to each in exchange. Of course, the first person encountered may not be amenable to a goods ratio acceptable to the first person, so it might take some hunting around, but with a small enough number of people within a tight enough spatial radius, the search can take only so long. Nonetheless, the search for the double coincidence of wants – remember that the person with the goods our searcher wants has to be agreeable to what our searcher has to

offer as well as to the quantity of it our searcher is willing to offer for a unit of what he has – costs real resources, time at the very least. But time can be used in production, resting (maintaining the labor force), courting, and consuming. Some of the transactions time has to come out of one of these other uses of time.

As the diversity of a society and its economy expands, the time and other resources devoted to barter transactions can be expected to rise. At some point there will be a cost margin at which it is cheaper to allocate part of one’s resources to a widely acceptable good in order to acquire goods that one wants but doesn’t produce. Use of such a good reduces the requirement for a double coincidence of wants to a single coincidence of wants: you only have to find somebody who wants your “widely acceptable good,” not your sack of beans or something that you don’t have. Of course, you still have to find somebody who is willing to part with his own goods for an acceptably (to you) small number of units of your widely acceptable good. Of course, this widely acceptable good is what we call money. You still have to tie up real resources in it, but fewer real resources (in value terms) than you would use in barter transactions. And the way the money ties up real resources may be far from obvious to individuals.

9.1.2 *Money as a store of value*

Anything that is used as a medium of exchange, a person must be able to keep around until she decides she wants to transact with it. Anything used as a transactions medium must be storable without losing its value.

Nonetheless, many items are capable of maintaining their value, or even increasing it, during storage or some other form of nonuse. Let’s call such items assets. Not all assets will be readily acceptable in transactions. The value of some assets is difficult to predict; if circulated at their current value, a person accepting some units of such an asset might find himself holding less value than he planned on. Liquidity is a characteristic of money as a store of value that distinguishes it from many other assets with storable value. In definitional terms, liquidity is the property such that an asset can be expected to

retain a very high proportion of its value – in the limit, its full value – in an immediate redemption. A house is a desirable store of value, but most houses can't be converted into their full value in an array of goods of the owner's choosing immediately. It could take several weeks, even months, to find a buyer willing to offer what the owner finds acceptable. In such transactions, it becomes apparent that the "full value" of the house actually isn't known until the transaction in which it changes hands for some particular value occurs. The high liquidity of money avoids this problem – or, depending on the particular money, reduces it.

9.1.3 *Money as a unit of account*

Measuring the value of a good in terms of another good is using the latter good as a numeraire. Measuring the values of many different goods in terms of a single good is using that numeraire good as a unit of account. Using the same good as a medium of exchange and a unit of account simplifies the calculation efforts involved in the medium's use, but it is not necessary that the same good provide both services. The use of the guinea as a unit of account in Britain long after it ceased to be used as a coin is just one example. In ancient Egypt, the use of the cow as a unit of account surely was not paralleled very often by its use as a medium of exchange. And besides, the concept of a cow as a unit of account presupposes some specific description of a cow, not just any one that happens along. The standard, unit-of-account cow might be a relatively ordinary cow in terms of its attributes, but many cows on the street or in the field would deviate from it in a number of characteristics. What would the characteristics of the standardized, unit-of-account cow include? Surely weight, undoubtedly health, probably age or an age range, possibly its track record in calving and in milk production. We quickly get the impression that the fewer dimensions on which a unit-of-account good might vary, the easier it would be to use.

9.1.4 *Stability of value*

Stability, or at the least, predictability, of value removes considerable risk from all the services

that a money provides. Using crops as a medium of exchange, as opposed to simply as a standardized unit of account, would introduce predictable seasonal fluctuations into the value of all transactions that were independent of the goods being exchanged. While predictability replaces stability, crops still are subject to less predictable value changes as a result of production shortfalls, and not necessarily even local ones if the crops are traded.

The traditional monetary metals – gold, silver, and copper or bronze – offer superior stability properties, but even with them, new strikes can rapidly augment a monetary stock and affect its value. If a society had a choice of these metals for its monetary standard, what factors might affect its decision? First, there is the matter of domestic, as opposed to foreign, supply. This might or might not be a major factor, but in ancient times foreign supplies would have involved considerable transportation costs, which could, of course, have been offset by use of a smaller stock and smaller coins (or dumps, in precoinage cases). Second, there is the matter of the diversity of supply. If one of the metals were mined in only one or two places, there would be greater possibility of supply instability and hence, value instability. A metal available from multiple sources would offer greater stability of value because the mining out of one or another source or new strikes at another source would affect a smaller proportion of the total supply.

9.1.5 *Monetization prior to currency*

It appears that in antiquity, throughout the Mediterranean societies, the separate services of money appeared at different times, with coinage being the last incarnation of the store-of-value and medium-of-exchange services. Stated alternatively, monetization occurred through the introduction and use of separate technologies, some hard (physical), some "soft" (intellectual). Hacksilber was used extensively in third-millennium Mesopotamia, providing both services (Snell 1995, 1490–1494), and the deben and kite were used as a unit of account system in Egypt centuries, if not millennia, prior to the introduction of any medium-of-exchange devices (Kemp 2006, 319). Kim (2002) stresses

the separation of coinage, a late innovation, from money generally. In Archaic Greece, he finds evidence for early introduction of small-denomination coins indicating early use by ordinary people. Kim observes that scholars looking for revolutionary social impacts of the introduction of coinage have found essentially none, implying that most of the consequences of monetization had occurred much earlier with the use of silver dumps and unit-of-account services.

9.2 The Types of Money

Money can be classified into two major types, commodity moneys and credit, or fiduciary, moneys. Figure 9.1 shows the subtypes of each major category. Looking at the left-hand side of that figure, it is worth pointing out that the world severed all its ties with commodity moneys only in 1973, so much of the recent (past hundred years or so) experience with money has considerable relevance to ancient moneys. Turning to the right-hand side, although paper money is a relatively recent innovation, banks and bank-created money have a long, even ancient, pedigree.¹

9.2.1 Commodity money

Commodity moneys can be either full-bodied or representative. The former include such forms as precious metal dumps, base metal objects such as the Archaic Greek obeloi or spits, minted coins, cowrie shells, stone coins (for example, jade), and so on. Representative forms of commodity moneys include paper money such as the recent U.S. silver certificates, which were terminated in 1967. These representative commodity moneys can be issued as fractional reserve moneys, in which more of the representations are issued than exist in the actual commodity money. Alternatively, a 100% reserve representative commodity money system keeps the number (value) of the representations equal to the value of the backing commodity. Either way, they are essentially warehouse receipts for the metal money they represent.

9.2.2 Credit money

The credit moneys may actually be more familiar to contemporary readers than the commodity moneys even though the commodity moneys have continued until quite recent times – and may eventually reappear. The three major categories

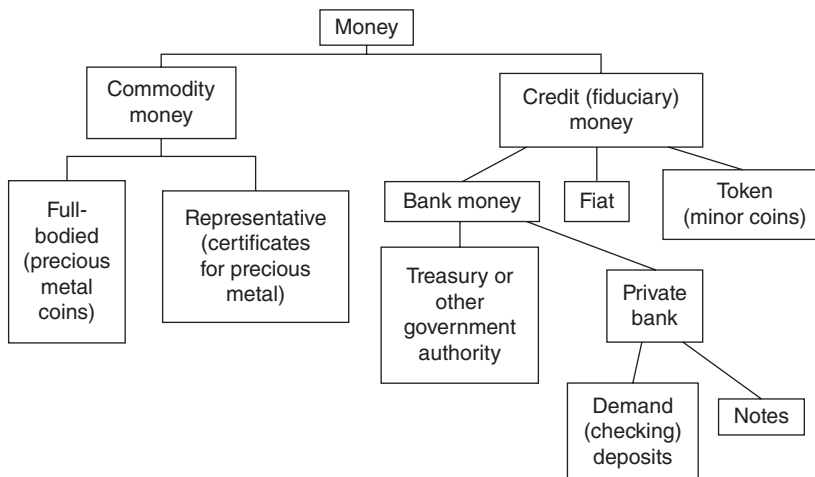


Figure 9.1 Types of money.

of credit moneys are tokens, fiat currency, and bank money. Examples of tokens are base-metal coins with intrinsic value substantially less than face value. Fiat currency in recent times is a paper money issued with no backing other than the public's confidence in the issuer, usually a government. An example of fiat currency is the paper money issued by the Continental Congress of the United States during the American Revolutionary War as payment for their purchases of goods and services. With this method of issue, its supply was not subject to any particular discipline, and its value quickly depreciated to the point where the expression "not worth a Continental [dollar]" became synonymous with worthlessness.

9.2.3 *One special case of credit money: bank money*

Bank money is the third major type of credit money. The banks issuing this type of money can have some official, governmental or quasi-governmental status, such as a central bank, or can be private. Central banks, such as the Federal Reserve System of the United States or the Bank of England in the United Kingdom, typically issue their own bank notes ("notes") as currency. Private banks, at various times and places, have issued their own bank notes which have circulated as currency. In the United States experience, while the notes may have had the same nominal denominations of the identical monetary standard (say, 1 U.S. dollar, or 5 U.S. dollars), they exchanged against one another at varying exchange rates, according to the community's assessment of the issuing banks' ability to redeem them on demand.

Of somewhat more importance to our particular interest in ancient monetary systems and behavior is private banks' creation of money through the establishment of checking and other deposits. The total value of checking deposits is included in the "narrow" (that is, the least inclusive) definition of the money supply in contemporary economies. Checks are accepted at face value in payment for purchases and are redeemable for reserves on presentation at the

bank issuing the check. Thus their liquidity is as high as that of currency or coinage. While checks are a relatively recent development among transactions technologies, the existence of deposits at banks in Classical and Hellenistic Greece, combined with loans made from the deposited funds, implies those banks' ability to create money beyond the drachmas minted by the government. Official regulation of private banks has varied considerably even over the past two hundred years of banking history in the United States, and contemporary levels and types of regulation are certainly not requisite for money creation by banks, although they do generally reduce risks of default and insolvency.

9.3 Some Preliminary Concepts

Several concepts are used so frequently in any discussion of money that it will be useful to introduce them properly and precisely. First, we have introduced the concept of the price level in Chapter 3 as one example of a price index. Rather than concentrate on the properties of index numbers as we did there, we will relate the price level to the value of money. Second, inflation is a very loosely used term in colloquial language, but it has a quite precise meaning in monetary theory. Third, the distinction between what are called "nominal" and "real" prices (or, more generally, values) is critical to distinguish in any consideration of behavior involving money.

9.3.1 *The price level*

Take the price of every good and service in the economy (p_i), measured in units of whatever numeraire, and multiply it by the quantity (number of units) of the respective goods and services (Q_i): $p_1Q_1 + p_2Q_2 + \cdots + p_nQ_n$. Next, add it all up and hang on to the resulting sum: $\sum_i p_iQ_i = Z$. Then divide each p_iQ_i combination (call these the expenditure on good or service i) by Z to get each good's and service's expenditure share: $p_iQ_i/Z = \gamma_i$ (of course, $\sum_i \gamma_i = 1$). Now, weight each good and service price by its expenditure share and add the expenditure-weighted

set of prices to get the price level: $\sum_i \gamma_i p_i = P$. This is the price level in this particular period.

When we proceed to construct the price level for the next period, we need concepts from the theory of index numbers to ensure that we're measuring the same things in different periods. Stated roughly, we will want to use the same set of expenditure shares in the different periods. This lets us compare the prices of the same bundle of goods and services over time; otherwise, we could be comparing the "bundled" price of an apple and two oranges with the bundled price of two apples and an orange. It'd be hard to decide what we were really comparing. However, if the quality of some of the goods changes over time, we would want to compensate for that by expanding the "quantity" of the good to represent the fact that a single unit of the new version offers a larger flow of services than a single unit of the older version. During ancient times, such quality corrections might or might not be particularly important as representations of technological improvement, but they could be significant if, say, we were measuring olive oil or wine arriving from different places and we suspected that they were of different qualities. We would want to weight a unit of the higher quality product more heavily than a unit of the inferior product. There are a lot of details here that we'll skip over; some of them you can retrieve from Chapter 3, others would require considerably more exposition.

Now that we've talked about constructing the price level for the second time period, we're in a position to talk specifically about what a "time series" of a price level variable would look like. The most common form of a price level variable extended over several time periods is as an index – the ratio of the current period's price level to the base-period's price level. The choice of the base period has a certain degree of arbitrariness to it, but as we saw in Chapter 3, that choice will have some effect on our assessment of the progress of the index over time.

Now, let's consider what determines the magnitude of the price level. We offer an introductory peek at a relationship we will consider at further length in section 9.4, the "quantity equation": $MV = PT$, in which M represents the quantity of

money, V is its velocity of circulation (that is, the number of times during the period in question that each unit of money turns over – changes hands – on average; we will discuss V much more closely in section 9.4, since it contains all this relationship's information on the demand for money), P is the price level (possibly in an index form), and T is the total number of transactions in the economy during the period. Let's explore what this relationship says. Suppose for the moment that V is fixed, for whatever reason, and that T is given by purely technological factors describing production possibilities. Then if the amount of money (say, the number of coins, all of the same denomination) is larger, the price level will be higher. The price level responds to the quantity of money, not vice versa. If V and M stay at the same levels but T gets larger, the price level will fall, because the same quantity of money, traveling at the same average velocity, has to cover more transactions. Conversely, if T were smaller but V and M remained constant – say, a war destroyed part of the productive capacity of the economy but left all the money intact – the price level would be higher.² So the price level can be altered by changing the quantity of money, the size of the economy, the velocity of circulation, or any combination of these changes.

We can rearrange the quantity equation to get a value insight into the interpretation of a price level: $P = MV/T$. The ratio M/T is, roughly, money over goods and services; its inverse is goods per unit of money, or the value of a unit of money in terms of the goods it will purchase. So, the price level is the inverse of the value of a unit of money. This is a flow concept, because the entire relationship $MV = PT$ describes what happens to money, goods, and prices *over* a particular period of time.

9.3.2 Inflation

Inflation is the rate of change of the price level over time.³ It is to be distinguished clearly from relative price changes of individual goods – say, a change in the price of dates relative to the price of lentils. Inflation is a purely monetary phenomenon, everywhere and always. Relative

price changes are real phenomena that may occur without the existence of money.

Although inflation is a monetary phenomenon, it may have “real” causes – that is, causes originating in physical production and consumption in the economy rather than changes in money. The causes of inflation are diverse. For instance, the inflation in U.S. currency (“Greenbacks”) during the American Civil War, from 1862 till about 1867, was caused by rapid expansion of a currency whose convertibility against gold had been suspended (Mitchell 1903). On the other side in that same war, the Confederacy experienced a considerable inflation, not because of excessive monetary issue, but largely because the territory in which the currency was accepted shrank after 1862: as the U.S. Army occupied more and more southern territory, a roughly constant quantity of money was left to chase a shrinking volume of goods and services (Lerner 1956). The Spanish importation of South American silver and gold in the sixteenth century was the principal cause of the ensuing rise of prices in Spain and subsequently in the rest of Europe (Hamilton 1934). Similarly, the worldwide price rise of the 1850s was attributable to the great gold discoveries in California and Australia in the late 1840s and early 1850s (Rockoff 1984, 619, 621 Table 14.1, 623). The European hyperinflations of the 1920s and immediately following World War II were caused by grossly excessive currency issue, injected into the economy by government purchases (Cagan 1956; Friedman and Schwartz 1963).

Deflation, of course, is just inflation with a minus sign – a decrease of the price level. Deflation also is a monetary phenomenon, with causes as diverse as those of inflation. The causes of the great, worldwide deflation of the 1930s are still being debated.

It is useful to distinguish inflation from just differences of price levels. Figure 9.2 shows the progress of a price level over a lengthy period of time. From time t_0 to time t_1 , the price level is constant; there is no inflation. At time t_1 , the price level takes an instantaneous jump, which could be called inflation, but since it takes place at virtually an instant, dividing the change in the

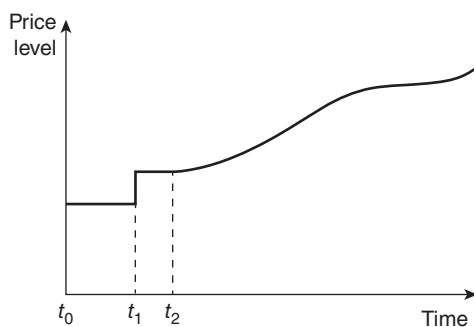


Figure 9.2 The time path of a price level.

price level by the length of time the change took to occur would give us an infinite inflation rate for a bare instant. We could debate the semantics of that change, but once the instant is over, the price level is higher than it was before time t_1 . From time t_1 to time t_2 the price level is again constant – there is no inflation, just as was the case over the time from t_0 to time t_1 . We cannot say that the period from t_1 to t_2 is “inflationary compared to the period from t_0 to t_1 ”; it just has a higher, and completely stable, price level. However, beginning at time t_2 , the price level begins a gradual ascent, which we *can* call inflationary. At all times after t_2 , we have nonzero inflation. (Does Figure 9.2 show any period of deflation?)

9.3.3 “Nominal” versus “real” distinctions

As a beginning to our explanation, “nominal” refers to prices or values denominated purely in monetary terms, although the use of money as a numeraire is only customary in dealing with contemporary issues, not absolutely necessary for the distinction. “Real” refers to a price or value denominated in the numeraire of a specific date. Thus, when you read the financial page of a newspaper you see that some price is referred to as being in “current dollars” or “current euros.” These are nominal prices. A few lines down on the page you may see what purport to be the same prices or values, but now with different numbers attached to them, and referred to as

being expressed in “constant (or, say, 1993) dollars.” These are “real” prices, although they are no more real in the phenomenological sense than the nominal prices. “Real” is strictly a technical term.

Why use the real-nominal distinction? The “real” version of a price or value is intended to make the purchasing power of a price or a numeraire of a good directly comparable across time periods. Suppose a bag of chickpeas costs 1 shekel this period. Two years later, we find that the same bag of chickpeas costs 3 shekels. It makes a lot of difference to us whether our daily earnings, which were, say, 4 shekels in the first year, are still 4 shekels or have gone up to 12, in proportion to the cost of the bag of chickpeas. It makes the difference between starvation and indifference. This is why, when confronted with, say, drachma prices of commodities at two well separated times in the Classical period in Greece, it’s difficult to determine what they mean, whether they are identical or different. If they are the same, unless we know what happened to the drachma price level over the period, we can’t say whether the price of the commodity remained the same or changed. Similarly if the prices are different: we don’t know whether the real value of the commodity changed or remained constant. To pretend that we can interpret the meaning of such prices without knowledge of the general price level is simply going beyond the evidence. Now, on the brighter side, if we have an array of other prices in each period, we can make some inferences from their proportions relative to the commodity in question. If they are all “about the same,” an admittedly crude measurement but possibly the best that we can do, then we are on firmer ground in suspecting that the general price level didn’t change much and consequently the observed drachma price in the second period is pretty (or reasonably) close to a real price even though, technically, it’s still a nominal price because we haven’t divided it by a price level deflator. Thus, real prices let us make observations on relative prices between time periods. Strictly speaking, we can’t make such observations if we have only nominal prices.

Dividing nominal prices of a particular date by the price level deflator of the corresponding date

lets us account for any inflation that may have occurred during the intervening period. Inflation also affects the value of money (the reason why we had to deflate the current prices by a price level index in the first place) and interest rates. Consider the effect on money first. If we observe that the money holdings of some individual at some point in time are, say, two talents, and two years later are, say, three talents, we don’t know how much effective money he really had unless we know how much stuff it would purchase. A simple observation of the number of talents is what is called “nominal balances” (money balances) – in the symbols of the quantity equation, just M . The value of nominal balances divided by the price level yields what is called “real balances,” M/P . The importance of the distinction between nominal and real money balances is that money is supplied in nominal terms but demanded in real terms. Nonetheless, each observation on nominal balances has a corresponding observation on real balances, even if we have difficulty distinguishing the difference empirically at the distance of some two millennia. We will discuss the behavioral implications of this distinction further in the next section.

We turn now to the distinction between nominal and real interest rates. Think about getting a 10% per year interest rate on your savings account. For each 100 dollars in your account, you’ll have 110 this time next year. But once you think about it a bit, what will 110 dollars buy next year? It depends on what happens to the price level between now and then. Only if the price level is perfectly stable – that is, there is a zero rate of inflation between now and then – will 110 dollars a year from now buy 10% more of the same bundle of goods it could buy now. What would the interest rate have to be to be equivalent to a 10% rate in the presence of ongoing and expected inflation? The answer to this question gives us the nominal interest rate. The nominal interest rate, just like a nominal price, is what is quoted. The real interest rate is the nominal rate minus the expected rate of inflation; in symbolic terms, $r = i - (\dot{P}/P)^e$, in which i is the nominal interest rate – the one we find quoted on clay tablets – and $\dot{P} = \Delta P / \Delta t$, or the time rate of

change of the price level, which is, of course, just the definition of the inflation rate; the superscript ε indicates that this is the expected inflation rate, not necessarily the actual rate once everything is said and done. People make most of their decisions on the basis of real interest rates rather than nominal rates. The difficulty for analysts is that the nominal rate is much easier to observe than the real rate; in fact the real rate is not directly observable but has to be inferred. Nonetheless, if as students, we insist on making predictions on the basis of nominal interest rates when the people we are studying are making their decisions on the basis of real interest rates, we should expect to produce poor predictions (or explanations, if we're "predicting after the fact").

9.3.4 *What people in antiquity knew*

It is a reasonable thing to ask at this point whether agents in the ancient economies understood and could measure these concepts that we claim are so important. Let's take these two questions in reverse order. Could they measure them? Undoubtedly not. We have difficulty measuring them today, although to give ourselves a break, we have a much more complex set of goods and services to measure than they did, and more of them too. The development of index numbers dates only from the late nineteenth century, and many problems in their construction have been identified only in the twentieth, and implementation technologies still lag behind theoretical understanding.

To rephrase the second question more productively, could they have identified inflation when they experienced it? A substantial enough inflation would be hard to miss, but it must be said that rapid inflations are easier with paper currencies than with pure commodity moneys. Stated alternatively, it's a lot harder to expand the gold or silver supply at 30% or 40% a year for several years than it is to leave the printing press on overnight. Inflation would have been more likely to be a longer term, chronic phenomenon in ancient economies on pure commodity money standards than a short-term, acute one, but evidence does exist that price levels changed

noticeably within what could be considered an individual's economic planning horizon.⁴ Temin (2006, 149) notes that the Roman banks disappeared during the course of the third century C.E., as the formerly slow pace of inflation increased to a gallop, leaving the *argentarii* with insufficient time to learn the difference between nominal and real interest rates to survive.

Schumpeter's assessment of the economic content of Plato's and Aristotle's writings identified a recognition by Aristotle that the value of silver and gold money was not immutable (Schumpeter 1954, 56, 62–63, esp. 62 n. 4), which is but half a step from identification of inflation. Nonetheless, such an approach to deciding whether or not the average run of people, in the conduct of their everyday affairs of staying alive and fed, recognized the phenomenon of price-level change has its risks, and is rather like looking to the contemporary works of, say, Wendell Berry, to find out what ordinary people actually do to stay warm in the winter. A brilliant philosopher's idea of the good life need have little relationship to how his or her contemporaries conduct their everyday lives.⁵ Far safer to depend on observations, however imperfect, of behavior in making inferences on implicit understandings which, while not reaching the attention of contemporary (that is, ancient) scientific thought, could have guided mundane behavior.

9.4 The Demand for Money

Monetary theory is composed of theories of the demand for money and the supply of money. Demand and supply are the backbone of economic theory, but little theoretical attention is given to the demand for and supply of specific goods other than money. This section opens with a discussion of why money is singled out for this special treatment yet is still believed to be subject to the same general laws of demand and supply as other goods. We follow this subsection with separate subsections on the two major approaches to monetary theory and a closing subsection on how these two approaches to money have been hammered together to form a largely

agreed upon paradigm for studying the influences of money.

Some readers might wonder why we don't just go straight to the last subsection without wandering through the wilderness of the contentions of a notoriously contentious group of scholars. Reasonable question. The rhetoric of the Keynesian revolution left a popular impression that there was little of value in the way of monetary theory prior to 1936. However, Keynesian monetary theory was built on the foundation of the neoclassical quantity theory, to which Keynes was himself a contributor. Keynes' monetary theory of the *General Theory* was an advance over the existing models in the quantity theory tradition, primarily for its casting of monetary theory in terms of capital theory. The neoclassical version of the quantity theory, with its focus on the transactions services of money, adopted a commodity view of money while keeping its services as an asset (store of value) in the background. Keynes treated the transactions demand for money relatively perfunctorily, being more interested in its role as an asset and its relationship to the interest rate. While the focus on the available models from the quantity theory as the world went into the Great Depression was on the long run – the period over which an economy fully adjusts to changes of various sorts – Keynes focused on the adjustment process itself – the immediate and short-run impacts of shocks to an economy. The short-run focus fostered the Keynesian assumption of a fixed price level, and eliminated, or at least blurred, the distinction between real and nominal interest rates, which formed an important mechanism for adjustment of an economy to changes in the money supply in the quantity theory. There was much of value in the neoclassical quantity theory, which Keynes indeed kept in his own new theory of money. Subsequent development of the quantity theory, particularly but certainly not exclusively by Milton Friedman and his students, produced a new hybrid paradigm in monetary theory that provided a vehicle in which disagreements could be pinpointed more closely to empirically researchable questions such as the sensitivity of the demand for particular types

of assets to their own returns and returns on competing assets.

9.4.1 *Measuring money*

If the only form that money took was, say, gold coins, measuring the quantity of money would be a straightforward task, if a laborious one. Just add up the number of coins of various denominations and take the grand-sum total. However, if gold coins *and* unminted gold dumps are both equally acceptable as money, we'd need to count the coins and weigh the dumps. But it's possible that the dumps would require weighing, whereas anyone can look at a coin and, recognizing its denomination immediately, determine how much money it is. Consequently dumps might be acceptable only by merchants (or others) who had scales available, and you might not trust just anybody's scales. So, while dumps are indeed part of the money stock, they may not always be as readily convertible into the goods one wants to buy as are coins.

This distinction between the “moneyness” of different types of assets provides the basis for different definitions of the money stock used in contemporary monetary analysis. The narrowest definition of money is called “M1,” and is composed of currency and demand deposits, because demand deposits (checking accounts) are as acceptable in payments as is cash. The next wider definition of money, called M2, includes M1 plus time deposits, which cannot be readily converted at their face value – you have to leave them in the bank for a designated minimum period or lose some of the interest you were getting on them. The next wider definition, M3, includes certificates of deposit and other so-called “near-moneys.”

Much of this may seem irrelevant to the ancient Mediterranean world, but we should not be so hasty in such a declaration. Athenian banks are known to have accepted deposits. If they issued statements of deposit that were acceptable in lieu of coins, then some measure of these “notes” or “checks” would have comprised part of the money supply, either of an M1-like variety or an M2-like variety: Cohen (1992, 12) contends

that: "...the deposit and credit mechanisms of [Athens'] *trapezai* constituted a ready source of 'token money' and thus an easy means of expanding its money supply" and expands on Athenian credit money in (2008, 81). And similarly in Rome: Andreau (1999, 132) reports that deposit bankers appeared in Rome between 318 and 310 B.C.E. and that the transfer of money without shifting material funds arose in the first half of the second century B.C.E., although no "endorsable checks or negotiable bills existed." It is possible that banks in other parts of the eastern Mediterranean world engaged in such practices at various times as well.

9.4.2 *The distinctiveness of the demand for money*

We have had a hint of the distinctiveness of money in the fact that the services of a large stock of money may be very much the same as those of a small stock of money. Additionally, it is widely agreed that money does not belong in the utility function but nonetheless a demand function for it exists. Then how do we derive the demand for money since demands for other goods are derived from the utility function? The logical structure of the answer to this question is clear: while money is not valued for the sake of its own consumption, it does lower the acquisition cost of the goods whose consumption *is* directly valued. So we should expect the demand for money to be based at one remove on the utility of the goods it can purchase.

The different services that money can provide – transactions services and asset storage services particularly – have provided different focal points for alternative theories of money demand. Theories based on the transactions services of money focus on the volume of transactions, or frequently, income, as the principal determinant of its demand, while theories that emphasize the asset role of money – that is, as a place to park one's wealth for a while – focus on the cost in terms of foregone interest earnings of keeping that part of one's wealth in a form that provides no direct return.⁶ However, many goods – both consumption goods and durables – possess multiple

characteristics and offer multiple services, yet their demands are formulated simply as functions of their prices and the consumer's income, with no attempt to distinguish between the demands for the separate attributes of the good. This attitude has characterized the synthetic approach to the demand for money.

The demand for money is the demand for a stock – a particular quantity at a particular time – but the transactions-based demand for money focuses on the stock required to support an expected turnover, or flow, of that money during a period. Consequently a comprehensive demand function for money must accommodate both of these aspects of the good.

Finally, and possibly most significant in distinguishing the demand for money from the demands for most other goods, is the fact that there are more complicated interactions between the supply of money and the demand for it than is the case with most goods. The aggregate demand for, say, sandals would be a function of the price of sandals, the prices of boots and socks (substitutes and complements) and possibly of tunics (feedback through the budget constraint), and income. The supply of sandals would also be a function of the price of sandals, and possibly of donkey and horse harness, the price of leather and leatherworking labor. The price of sandals is endogenous to the equilibration of demand and supply but the other variables in the demand and supply functions are unique to those functions. Thus, a shift in the price of leather, which appears in the sandal supply function, would have no effect on the demand for sandals, although it would have an effect on the equilibrium price of sandals because it shifted the supply curve. Similarly, a change in the price of socks (due to a change in the price of wool?) would shift the demand curve for sandals but would have no effect on the sandal supply function. In the case of money, some of the key variables in the supply function also appear in the demand function or affect variables that appear in the demand function, so it is not possible to consider independent changes in supply and demand. This fact makes it difficult to study monetary theory independently of what is currently called macroeconomics, or

the theory of aggregate employment, interest rates and the price level: frequently, the effect of a shift in the supply of money depends on how that change affects one or more of the determinants of money demand. This probably sounds rather abstract, particularly since we have kept under wraps the variables in the money demand and supply functions, but we hope this warning will illuminate some of the intricacies and potential analytical pitfalls as we turn to the theory of money demand, and, in the subsequent section, to money supply.

Having just noted the difficulty of studying the demand and supply of money separately, it may seem peculiar that we are following such a structure ourselves. Some of the monetary theories we treat below are more than theories of the demand for money, some of them being also theories of the price level and theories of the interest rate, and some of them predicting that how monetary changes affect the economy depends on how the money is supplied. This is simply unavoidable, but the independent treatment of the demand for money is still useful.

9.4.3 *Monetary theory and macroeconomics for ancient economies?!*

Probably more critical to the purposes of this book is the relevance of monetary theory at all, particularly with its close links to macroeconomics, to the ancient economies. This is a subject that has been more subject to pronouncements than careful consideration. I will not pretend to conduct original research on the question of applicability myself here, but I must offer some reflections on the possibilities. I have already noted the greater difficulty in rapidly expanding the money supply with a pure commodity money than contemporary and recent economies have experienced with various types of paper money. This is only one possible source of shock to an economy that could have repercussions on the demand for money, the price level, the level and structure of interest rates, employment, output, and consumption.

I suspect that employment may be the key element in this panoply. What reason do we have to believe that there was sufficient work done for hire by others in these economies to make the possibility of involuntary unemployment in the contemporary sense a reasonable one? Our attention is attracted in particular to the large urban centers of the ancient Near East, to Classical and Hellenistic Athens, and to Imperial Rome. Undoubtedly some of the residents of these cities were independent farmers, but the sizes of many of the Near Eastern cities suggest that a lot of farmers would have had to walk quite a way to their fields every day, which altogether reduces much of the motivation for living in a city in the first place. I suspect that a high proportion of the populations of these cities worked in what today would be called light manufacturing and services: making tools, making shoes, repairing shoes, preparing food for people who either didn't go home for lunch or supper (or were already up and about before breakfast) or didn't have any cooking facilities where they lived ("boarding houses"), donkey driving and other transportation and stevedore activities, pawnbroking, hauling sewage – just think of the thousand and one things that get done to permit people to live in large concentrations. Classical Athens is believed to have been full of slaves. Slaves have to be fed; food costs money (or whatever else one might exchange for it); the production of goods or services underlies the acquisition of money; even if slaves were held just for the prestige of having them, they still cost their upkeep (if a person thinks the slaves weren't put to work to provide their own upkeep – not to mention offering their owner a return on his or her investment – such a person needs to come up with another source for the upkeep).

Now, if something happens to change people's desires to spend, people who make shoes will have fewer customers, people who repair shoes may have more or fewer (think of increased maintenance substituting for new investment), pawnbrokers may find themselves keeping more junk they don't want and can't get rid of, donkey drivers and stevedores may find themselves with less to do. Whether the shock responsible for

such a chain of events is a change in the money supply, a change in one of the arguments in the money demand function, a tax change, a weather shock to local agricultural production, a change in export demand, or something else, the change in spending patterns will affect how much work nonfarm workers find themselves doing. Think of the interruption of payments to the royal tomb workmen at Deir el-Medineh in New Kingdom Egypt (Lesko 1994, 38; Brewer and Teeter 1999, 49). Called by any other name, this change in the amount of work people are able to find is a change in employment, or conversely, a change in involuntary unemployment.

Altogether, the linkages of the demand for money to other contemporarily defined macroeconomic variables may be worth not dismissing out of hand when thinking of the ancient economies of the Mediterranean and Aegean region. We've talked about talking about monetary theories long enough. Let's turn to them.

9.4.4 *The neoclassical quantity theory*

The quantity theory is an approach to monetary theory rather than a single model. Many different models can be constructed within this paradigm. They can ask different questions and give slightly different answers to the same questions simply because they frame the questions differently. The neoclassical quantity theory is essentially a theory of the price level, in which the supply of money, with a number of other things held constant, determines the price level. We introduced the equation of exchange in section 9.1 to help discuss the price level: $MV = PT$, or using aggregate income instead of transactions, $MV = PY$, where Y is aggregate income in the economy. The total volume of transactions will include many intermediate transactions that are excluded from final income, so the magnitude of what is called "income velocity" will be smaller than the "transactions velocity" associated with the variable T . As we noted earlier, the relationship predicts that a larger stock of money will produce a higher price level; this is a causal relationship.

Rearrange the equation of exchange to obtain a money demand relationship: $M/P = Y/V$. A

larger real income, which is determined by a production function, will increase the demand for real money balances, presuming the velocity is stable. What is meant by a stable velocity has caused problems. A stable velocity is not one that is fixed and unchanging, but one that is a stable relationship of a few variables. When those determinants change, velocity will change accordingly. This remained implicit in applications of the quantity theory early in this century and contributed to the deficiencies that prompted Keynes' new approach to money in the mid-1930s. Nevertheless, the quantity theorists of the early twentieth century, such as Alfred Marshall, A.C. (Arthur Cecil) Pigou, Dennis Robertson, and Irving Fisher recognized the influence of the interest rate on the incentive to hold cash balances as well as the role of uncertainty about future events, both price-level uncertainty and interest-rate uncertainty, the latter being particularly important in Keynes' new theory. The higher the interest rate, the larger the foregone interest income a person must accept to hold non-interest-bearing cash; consequently, the demand for real balances is negatively related to "the" interest rate (recognizing that there are lots of different interest rates). Inflation erodes the purchasing power of money, so a higher expected inflation rate in the near future will depress the demand for real money balances. The demand for money derived from the equation of exchange is the quantity of money required to support the transactions associated with income Y ; while this quantity is a stock, it is very close to a flow demand for money.

The principal alternative model of money within the neoclassical quantity theory tradition is called the Cambridge cash balance, or Cambridge k , approach. The relationship is specified as $M = kPY$, in which k specifies the fraction of income which people will want to hold as money balances. The equation can be rearranged to give a demand for real balances: $M/P = kY$. The proportionality factor, k , is equivalent to the inverse of velocity, or $1/V$. The formal appearance of the Cambridge formulation is identical to the equation of exchange but it asks a different question: given that money is used to make

transactions, how much of it would an individual, or the entire group of individuals in an economy, like to hold? Once we ask the question, "How much cash does someone *want* to hold?" we recognize that wants will typically exceed capabilities, and that in this approach people will have to decide among competing assets, sacrificing some interest income for the convenience of being able to conduct transactions.

Expressed as a money demand function, V and k can be written as functions rather than scalar (that is, single-number) parameters. We could write $V = V(r, W, \dot{P}^e/P)$, in which r is "the" real interest rate, W is wealth (some original neoclassical theorists would have been inclined to use income; it wasn't particularly well thought through), and $\dot{P}^e/P$ is the expected inflation rate. A lower velocity is the same thing as a greater demand for money. A higher interest rate increases velocity (decreases the demand for money – makes people want to spend their money more rapidly – turn it over faster). Greater wealth increases the demand for money, which lowers velocity. A higher expected inflation rate encourages people to get rid of money, which means spending it for goods and other assets; otherwise, their money holdings will fall in real value as the price level rises – "use it or lose it."

Even if there were no alternative assets that offered positive returns (nothing that offered an interest rate r), there would still be a cost to holding money – the actual (not the expected) inflation rate. (If people knew beforehand what the actual inflation rate would be, the actual rate would replace the expected rate in the money demand function; the expected rate is used in that function because it generally is the best people can do as far as forecasting goes.) Of course, if the price level is falling, money offers a positive return. As we noted above, the flow value of money is the inverse of the price level ($1/P$); the marginal stock value of money is "the" interest rate, or the percentage difference in the rate of change in the flow values of money between periods.⁷ This distinction is equivalent to the distinction we made for capital between the rental on a unit of capital used during a period of time

and the interest rate on some quantity of capital at a given moment in time.

The quantity theory paradigm offers two principal mechanisms by which an increase in the supply of nominal money causes the price level to increase, thereby bringing the level of real balances demanded into equality with the nominal balances supplied. The first is an expenditure mechanism, by which people divest themselves of money by spending their excess holdings on both goods and assets. They initially find themselves with more cash than they wish to hold, given their income and the current level of prices. Spending their excess balances drives up the price level because they can spend their newly excess balances faster than the economy (implicitly assumed to be operating at full employment, and thus full productive capacity) can produce new ones. The increase in the price level reduces their real balances. The second mechanism operates initially through interest rates. Banks find themselves with more cash than they need to hold for sound lending practices, so they lend out more. To place the increased volume of loans they have to offer a lower interest rate. The increased borrowing finances increased investment, which together with other increases in spending, pushes up the price level. The rising price level depresses the nominal interest rate below the real rate, and for a while the bank lending rate has to rise faster than the price level to bring the real and nominal interest rates into equality, the former of which reduces desired real balances and the latter actual real balances. Money is supplied in nominal terms but the demand for money is for real balances; the change of the price level via spending money more rapidly (operating on V or k), or the change in the interest rate via the credit-market mechanism, or a combination of both, operate to bring real balances demanded into line with the nominal supply of money.

One prominent proposition of the quantity theory has been the prediction of a change in the price level proportional to the change in the money supply in the long run. This is called the neutrality of money. Considerable theoretical research in the past several decades, as well as

theoretical analysis prior to Keynes' *General Theory*, has shown that exact neutrality is a special case. The implication of non-neutrality is that expansion of the money supply has effects on the real economy – real output, employment, and the interest rate.

While alternative theories of the demand for money are available and offer more flexible models, the quantity equation remains a useful framework for organizing our thought about money supply, money demand, aggregate income, and the price level. Thus, we should expect an increase in the price level if the supply of money increases and no real changes affect production possibilities in the economy. Alternatively, if productive capacity is growing but the money stock is not, the price level will decline as long as money demand (velocity) remains constant. These are long-run tendencies – that is, relevant to the time period during which the elements in the economy that can adjust have had time to move to their new, equilibrium levels.⁸ As we will see below, Keynesian monetary theory (and Keynesian macroeconomics in general) focuses more on the short-run transition paths between long-run equilibrium positions than on long-run equilibrium.

9.4.5 Keynesian monetary theory

Keynes' approach to monetary theory separated the demand for money into distinct components corresponding to some of the traditionally accepted services of money. Accordingly, he retained the transactions demand emphasized by the neoclassical quantity theory, which he made a simple function of income: $M_T^d/P = kY$, which is the same as the Cambridge equation (after all, Keynes was at Cambridge). Closely related to the transactions demand was what he called the precautionary demand, the quantity of money an agent will want to hold to avoid cash shortfalls in the face of unanticipated expenditures. This component of demand he also specified to be a function of income, and it has been common to simply lump these two components of money demand together since they are functions of the same variable.

The feature of Keynesian monetary theory that distinguished it from the neoclassical quantity theory was that, with the speculative component of demand, it viewed money from the perspective of capital theory. The principal innovation of the Keynesian approach to money was to emphasize the role of money as an asset as the basis of one component (and possibly a quite important one) of the demand for it. The speculative demand, Keynes specified as $M_S^d/P = \ell(i)W$, which you can read as "a function ℓ of i , times W ," not " ℓ times i times W ," where $\ell(i)$ is the asset price function. As the interest rate rises, the cost of holding noninterest-bearing money gets higher, and people will want to put more of their assets in forms that offer them a return. (If money happens to bear interest, which it generally does in checking deposits today, then the relevant variable is the difference between the interest rate on demand deposits and the returns available on other assets.) Thus $\Delta\ell/\Delta i < 0$. This functional relationship is multiplied by wealth, since wealth, rather than income, will be the relevant base against which speculative holdings will be determined. That is, harkening back to the interpretation of the Cambridge equation in the neoclassical quantity theory, the speculative demand for money predicts the proportion of a person's wealth he or she wants to hold in cash.

Once we view money as an asset – a store of value – we need to look around at the substitute stores of value. The principal ones, in any society, are bonds – essentially, loans that can be sold pretty easily – and real capital. Real capital can be held (owned) through direct ownership of assets like buildings, inventories, animals, slaves, equipment, vehicles and vessels, and so on. Participation in a joint trading venture, through, say, the supply of animals used for transportation, advancing of money for purchase of provisions, loans to purchase inventories sent out on the venture, and so on, qualify for the condition we've called "holding one's assets in real capital." Joint ownership of a vessel – or sole ownership of one – would qualify too, of course. Accomplishing such asset holding through a stock market makes selling one's part in such a venture easier

but the act of putting part of one's assets into real capital is the same whether one does it through a stock market or through one's friends. A bond is a loan whose contract promises to pay a fixed amount each period (typically a year) plus the amount of the original loan at some specified date in the future. A bond market makes the issuance of such loans easier because whoever buys one of them can unload it pretty much when he or she wants, although the price can vary rather unpredictably – one can make either capital gains or capital losses on the sale. Thus, altogether, we believe that the absence of organized stock and bond markets from the ancient Near East and Aegean does not eliminate the applicability of Keynesian monetary theory to ancient monetary behavior, although one may need to use a bit more ingenuity in its application than there than to contemporary economies. Consequently, the extensions of Keynes' original specification of the asset demand for money (a more modern term than speculative demand) typically contain several different interest rates (or rates of return).

The Keynesian money demand function is commonly written as $M^d/P = kY + \ell(\mathbf{i})W$, where the bold "i" indicates an array of interest rates. Two of Keynes' most important objections to neoclassical monetary theory of his time were (i) its presumptions of flexible prices (and hence of a flexible price level), which in turn (ii) led to full employment. One of the hallmarks of Keynes' new model of the aggregate economy was its assumption that the price level was fixed, which permitted unemployment but also confined the model's principal applicability to some short run (of undefined length). This fixed-price-level assumption let Keynes use the nominal interest rate, which we represent here by "*i*," ignoring the distinction between nominal and real interest rates we introduced above and which had been prominent in the less formal treatments of the neoclassical quantity theory. This point remained contentious between Keynesians and quantity theorists for a number of years, but the importance of real interest rates to the demand for money has been generally accepted in recent decades.

According to Keynes' theory of how "the" interest rate affected the demand for money, investors looked at the current level of the interest rate relative to what they expected to be its "normal" level. If the current rate were above the normal rate, people would expect the rate to fall, and referring back to our formula for the relation between the price of a capital asset and the interest rate, this would cause a rise in bond prices (Keynes restricted the alternative assets he considered to bonds; not a necessity). Expecting bond prices to rise, people holding cash would rush to exchange their cash for bonds before the price rose. Conversely when the interest rate was considered to be below normal: everybody would want to hold cash so as to avoid the capital losses imposed by a rising interest rate. The role of the interest rate in the Keynesian theory differs from the role assigned in the transactions-based, neoclassical quantity theory, where the interest rate reflected the opportunity cost of keeping assets in noninterest-bearing cash instead of in a form that would yield a positive return.

This attitude toward the role of the interest rate in determining the demand for money balances was the basis of what Keynes called the "liquidity trap": at some low interest rate, the monetary authority could issue an unlimited quantity of money without lowering the interest rate further, because everybody would be expecting the rate to rise and would be willing to hold all the money they could get. Keynes himself doubted the empirical importance of such a situation, but it has characterized the Japanese economy from the early 1990s through the present, and has been expressed as a concern in countries' sluggish recoveries, to the extent recoveries have occurred, from the Great Recession of 2008 to the present. However, as you probably noticed, this view of the asset demand for money being determined by the relationship of the current interest rate to some level other than "normal" offers considerable "play," because the level that investors consider normal at one time might not be considered normal at another time. Consequently, Keynes himself considered the demand for money to be

highly unstable and consequently an unreliable instrument of government economic policy, at least as far as stimulating employment is concerned. This Keynesian feature provided a clear focal point for the empirical research on the stability of the money demand function that characterized subsequent research on monetary theory.

Since the overall Keynesian model that contained Keynes' new money demand function had a fixed price level, the quantity of money in the economy could not determine the price level. Instead, it determined the interest rate. Nevertheless, it would be legitimate to use a Keynesian money demand function in an aggregate model with a flexible price level.

9.4.6 *The contemporary synthesis*

While differences of opinion still exist regarding the most useful approach to viewing money, it is reasonable to say that a substantial synthesis of the insights from the neoclassical quantity theory and the Keynesian theory has been accomplished. Possibly the most important outstanding difference is the portfolio view of the demand for money, associated with the Keynesian tradition (Tobin 1958), and the aggregate demand for money function that Milton Friedman called the new quantity theory but is in effect a combination of Keynesian and neoclassical quantity theories (Friedman 1956). Since we have already treated the portfolio model in Chapter 7, we will concentrate here on the modern quantity theory and a variant on the transactions demand for money.

Friedman's modern quantity theory casts the demand for money in the mold of ordinary demand theory, in which the quantity of a good demanded is a function of its own price, the prices of substitutes and complements, and income. Just as the demand function for goods is derived from the first-order conditions of a utility-maximization problem, a money demand function can be derived from a utility-maximization problem. I won't subject you to the details here, but we will talk a bit about the structure of the problem. I noted early in the

chapter that it is generally – universally – agreed that money is demanded not for itself but for what it can purchase. Consequently, some people have been reluctant to put money into the utility function along with other goods and services to derive the demand function for money, but other students of the issue have done that. It is not necessary, however, to put money into the utility function to obtain a money demand function from a utility maximization problem. Placing money into the set of constraints, subject to which the utility is maximized, will yield a demand function for money as a function of the prices of goods (the purchase of which, after all, is the ultimate reason for holding money), returns on alternative assets, and characteristics of the costs of conducting transactions and holding stocks of goods. The same problem also yields demand functions for the other assets available and the goods in the utility function (with only the goods, none of the other assets, in the utility function).⁹ Some versions of this approach to the money demand function have emphasized the wage rate as an indicator of transactions costs, on the grounds that labor time is the principal component of transaction costs (Dutton and Gramm 1973; Karni 1973; Dowd 1990; McCallum and Goodfriend 1987, 777). These models, which have found empirical support, indicate that higher wages increase the demand for money, independently of the effect of the wage rate on income or wealth. Individuals and societies with more valuable time are willing to hold more money to conduct their transactions, for any given opportunity cost of holding money, than are individuals and societies with lower values of time.

Having offered some reassurance that Friedman's money demand function can be derived from the first principles of individual economic behavior (I'm glossing over aggregation problems entirely for the moment), we can present the functional form he developed. His demand for real money balances is a function of the interest rate on bonds, less the expected change in the bond interest rate; the real rate of return on equities, less the expected change in that return; the expected rate of inflation (the rate of return

on holding stocks of goods); and wealth. Some economists would modify this specification to put the price level in the function, either in lieu of dividing nominal balances by it, or in addition to that, on the grounds that the demand for real balances may not be entirely neutral to the price level. In symbolic terms, this money demand function can be written as $M^d/P = f(r_b - i_b^e/r_b, r_e - i_e^e/r_e + \dot{P}^e/P, \dot{P}^e/P, W)$, in which the subscript *b* indicates the bond interest rate, superscripts *e* indicate expected rates of change (the rates of change denoted by the dot over the variable), and *W* is wealth. Sometimes a measure of permanent income is used in place of wealth. Increases in each of the interest rates and the expected inflation rate depress the demand for real balances. This formulation of how the various interest rates on alternative assets affect the demand for money remedies the instability problem that the Keynesian speculative demand introduced and combines the opportunity cost aspect of the real interest rate from the neoclassical quantity theory with the potential capital gain or loss emphasized in the Keynesian model.

Empirical research on the demand for money in a variety of countries and over periods ranging from the nineteenth century to late in the twentieth generally has yielded evidence for negative interest elasticities, positive wealth and income elasticities, and negative elasticities with respect to expected inflation. Neutrality with respect to the price level has been a more contentious issue, but theoretically, under many conditions, the quantity of money can be expected to have long-run effects on such real variables as employment and income. In other words, in many circumstances money can be more than just a "veil over the real economy."

Another model of the transaction demand for money, developed by Baumol (1952), offers the interesting implication of increasing returns in money holdings. We will not develop the final money demand formulation from the model's first principles, but the objective of the agent demanding money is to minimize the costs of holding cash to meet regularly scheduled payments. The alternative to holding cash is to put

the cash into bonds that yield the interest rate *r*, but there is a transaction cost, *b*, of switching between cash and bonds, regardless of the size of the transaction. The least-cost quantity of money to hold to meet these known payments is $M^d/P = 1/2\sqrt{2bT/r}$, in which *T* is the volume of transactions per period. The increasing return arises from the fact that when the transaction volume rises by, say, one unit, the quantity of money held will rise only in proportion to the square root, not in a one-for-one relationship. There appears to be some empirical support for this type of relationship between income (not wealth, or permanent income), representing transactions, and money demand. If this effect were particularly strong at the individual level in a society, greater disparity in the personal distribution of income would depress the aggregate demand for money.

9.5 The Supply of Money

Since money is a good, whether we are referring to a commodity money specifically or to some variety of fiat money, it has both a supply function and a production function. Commodity moneys, such as gold and silver, are produced by mining and minting if we trace the commodity back to its source. Nevertheless, in some cases it makes sense to consider the supply of money as coming from exports, in which case we need not concern ourselves with going all the way back to the production function for money and in fact may not need a very complicated supply function: domestic residents just offer goods for money that comes from outside the country or economy. However, if it makes sense for the analysis of some particular problem to consider the agent who issues the money in question, we would be well served to consider that agent's supply function.

Two industries in particular are involved in the supply of money – the mining industry and the banking industry. Various agencies of government also get into the act in one way or another, although they aren't necessary for the existence of

money. Sometimes these agencies are little more than an intermediary between the producers of money and the holders of it, as could be said of the role of a mint; sometimes they are just extra banks in the system. We won't get into the engineering details of mining here, but some aspects of the economics of mining will be important in the analysis of commodity moneys. If banks are technologically equipped and legally permitted to engage in fractional reserve lending, they create money. If not, they still are in the financial intermediation industry, which helps borrowers find lenders and vice versa.

Before considering the contemporary theory, it is worth noting that the banking and financial systems of a number of ancient civilizations have been studied extensively. Beginning with the Ur III and Old Babylonian periods in the ancient Near East, De Graef (2008) has studied loan texts from two excavations for which archaeological contexts are available, testing what she calls the "traditional" theory about those texts, that the texts were broken or destroyed when the loan was repaid, which would imply texts documenting an enormous preponderance of unrepaid loans from those periods. Based on the stratigraphy and genre of the texts from the archive of Igi-buni from the Ur III period at Susa, and on the contents and context of the texts from Ur-Utu's archive at Sippar from the Old Babylonian period, she concludes that both sets of texts were most likely discarded postrepayment and in fact represented the remains of what today would be called "for-profit" business.

Moving to the Greek world (of primarily Athens, for want of evidence from elsewhere) some 1500 years or more later, Cohen (1992; 2008) and Millett (1991) present contrasting views of banking and financial intermediation during the Classical period. Cohen sees Athenian financial operations with structural parallels to many contemporary banking practices, while Millett sees lending and borrowing as primarily a way of cementing friendships among Athenian citizens, with professional bankers and lenders remaining a marginal business lending primarily to metics and people who otherwise got themselves into financial problems.

The Roman world has provided considerably more evidence to work with than has the Greek, although scholars of the subject typically remind their audiences of how much remains unknown. The development and operations of banking and various forms of financial intermediation have been presented in varying degrees of detail (considering that they cover some 600 years or more) by, *inter alios*, Andraeu (1999), Verboven (2009) and the papers in the Bogaert festschrift edited by Verboven, Vandorpe and Chankowski (Verboven *et al.* 2008), an admittedly nonrepresentative, if generally recent, sample. In the latter volume, Rathbone and Temin (2008, 371) conclude that financial intermediation was "more extensively available and flexible" in first-century C.E. Rome than in eighteenth-century England and that, "It [the Roman economy] was also an economy permeated by credit created through paper transactions, multiplying many times the level of monetization achievable through coinage alone." The financial services they identify include money changing, deposit accounts, making mandated payments, transfers between accounts, credit for auctions, loans to clients and third parties, guarantees for contracts and legal appearances, and in the provinces tax payments, and maybe "other as yet unattested services" (416), an array to which Verboven (2008, 226–227) would concur and add registering debts, transferring debts between parties, and standing surety. In subsequent work Verboven (2009, 95) has emphasized what he characterizes as functional monetary "modes" in Rome during the Late Republic and the Early Empire: currency, commodity, and, the most sophisticated, the account mode, which is a registration system for debits and credits employed by private individuals, nonbank firms, and various types of financial intermediaries. He offers extensive discussion of the various forms of operations falling under the account mode, including multiple-party as well as bilateral transactions (123–133). While there is fairly detailed and nuanced understanding of the institutions and operations of these financial systems, there seems to be scope to connect the activities of this ancient industry with the wider economy.

9.5.1 *Supply of a commodity money*¹⁰

The world's monetary systems have been entirely free of ties to some commodity since only 1973, so there is plenty of experience with commodity money systems such as those that existed in antiquity. While various animals and bolts of cloth may have provided some monetary services at certain times and places within our purview, the most accessible instances of money use – in the sense of availability of evidence – involve metals – gold, silver, bronze, copper, iron. Consequently we will discuss the operation of a commodity monetary system with reference to a gold standard: that is, whatever serves as the circulating medium is convertible into gold, generally at a ratio specified in advance. Under a gold standard, if the value of the monetary gold stock is equal to the total money supply, we have what is called a pure gold standard. (Not all gold in the economy or society in general need be devoted to monetary use; for example, some could be used as jewelry, paint, and so forth.) If the money supply is greater than the value of the monetary gold stock, we have a fractional gold standard (the remainder being made up of notes or other tokens redeemable in gold). If more than one metal is in concurrent use as money, the system is called bimetallism, even if more than two metals are involved; the principles are the same with two or more metals. A pure gold standard can be either a gold specie standard, in which gold coins are circulated, or a gold bullion standard, in which notes or other tokens are convertible to gold ingots with 100% gold reserves (that is, the volume of the notes or other tokens can't exceed the value of the gold ingots or we'd have a fractional gold standard).

In this section we will address the pure gold specie standard on the grounds that it is the template of most ancient monetary standards, but we should keep in mind three caveats: First, there may have been a number of instances in which banks operated on less than 100% metal reserves (that is, they made loans in excess of the value of their deposits, thus creating money in excess of the value of the metal stock); in this case we would have a fractional gold standard. We will

discuss the creation of money by banks in the following subsection. Second, there were a number of instances of bimetallism; we will discuss bimetallism later in this section. And third, the use of metal dumps rather than minted coins raises an issue not frequently dealt with in contemporary monetary economics, but certainly one that is not outside the capacity of contemporary theory to address. The use of dumps is structurally close to a system of private (that is, nongovernmental) issue of money. The transactions costs of weighing dumps with scales such as have proven relatively common among artifactual finds are equivalent to the costs of establishing and maintaining the variable exchange rates between the media of the various issuers such as the private bank-note era of the United States in the first four decades of the nineteenth century.

Two institutional rules are fundamental to a gold standard. First, the issuer must give a technical definition to its currency – specify the material, the weights of various units (coins), and the price of the material in terms of some unit of account. For example, a U.S. “eagle” coin weighed 232.20 grains of gold, 9/10 fine, and was defined as worth 10 dollars. Second are what could be called the monetization rules. Anyone could either obtain from the mint (either government or private) as much currency as they wanted as long as they provided or paid for the supplies of the raw material required and the minting cost; or had their coins melted down and returned in the form of bullion, again for a relatively small charge. This procedure is called free coinage.

Money is gold (but not all gold is money), and there are two sources of gold. It can be mined or it can come from overseas. If mined domestically, the resources (labor and equipment) used in mining are diverted from other production activities which, for simplicity, we can call commodities. If gold is imported, it is paid for with exports of some kinds of goods, and the resources devoted to producing those exports correspondingly must be diverted from the production of other commodities that could be consumed domestically. The money that is used is subject to a kind of user cost in the form of physical losses of gold

through abrasion and occasional losses of coins. In thinking about the supply of money in an economy like this, it is useful to consider that all the economy's resources are used in either commodity production or gold mining. You can think in terms of a production possibilities frontier such as we developed in section 13 of Chapter 2 (production in an entire economy) and used in Chapter 5: it describes how much an economy could produce of one good if it produces a certain amount of "all other goods" – subject to the total availability of productive resources in the economy.

The efficient production of gold, if it is supplied through domestic mining, is determined by the equalization of the ratio of marginal cost of commodity production to the marginal cost of gold production to the ratio of the commodity price to the gold price. We haven't said where the "gold price" comes from yet; we'll get there soon. Aggregate income in this economy consists of the incomes from commodity production and gold production, and the aggregate demand for commodities is determined by aggregate income and relative prices (this is pretty much like the demand curves discussed in Chapter 3). Of course, with the level of simplification (aggregation) we're using here, the only other price available to form a relative price is the price of gold, so the relative price we're referring to is the ratio of the commodity price to the gold price. The user cost of monetary gold also will affect the aggregate commodity demand curve because it describes the aggregate losses across the economy from using gold money in transactions for commodities.

The previous two paragraphs have described four relationships – production possibilities in the economy, the efficient allocation of resources to commodity production and gold mining, the resulting aggregate income, and the demand for commodities. These four relationships determine total gold production, total commodity consumption (or production; it comes to the same thing), aggregate income, and the relative price of commodities and gold. We still don't know either the price of gold or what could be called the price level, which is just the absolute price

of commodities (measured in monetary gold). We also don't know how much of the gold that is produced will go into the money stock and how much will be devoted to nonmonetary uses; nor do we know the size or value of the total money stock.

Once the government (assuming that it is the government that is issuing/minting the money) has decided on the price it wants to assign to gold (to use our earlier example, whether the 232.20 grains of gold in the eagle is going to be called 5 dollars, 10 dollars, or 1 deben), the absolute price of commodities, and hence the commodity value of money, are known from the cost conditions in production and the demand for commodities. Setting the price of gold in terms of the standard's numeraire also determines the demand for real money balances measured in terms of gold. The demand for money, which we haven't relied upon yet, will be a function of aggregate income, the relative price of commodities and money, and the user cost of money. The demand for money will be larger with larger aggregate income and will be smaller with a higher user cost of money. The relative price of commodities and gold is not a variable that our previous discussion of either the quantity theory or the Keynesian theory called forth. The relative price of commodities and gold is in this demand for gold money because it explicitly characterizes the opportunity cost of adding to the money stock: as this relative price goes up, newly mined gold gets cheaper relative to commodities and it is less costly in terms of commodities to add to the money stock. Consequently, the demand for money increases as the relative price of commodities and gold rises.

Now that we know the quantity of gold produced and the stock demand for money, we are able to determine how much newly mined gold goes into the money stock and how much goes to nonmonetary uses. Each period the fraction of monetary gold we identified as the user cost disappears, and to maintain the equilibrium size of the money stock, this much gold will go into monetary use to replace what disappeared through use. The remainder of new gold production will go into nonmonetary uses. The adding-up condition for the aggregate economy (recall Walras' Law

from Chapter 5) tells us that if we know the equilibrium quantities of $n-1$ of the goods demanded in an economy, we automatically know the quantity demanded of the n^{th} good – in this case, nonmonetary gold.

If, for some reason, the demand for nonmonetary gold increased, this change would set off the following set of reactions. First, at an unchanged commodity/gold price ratio, gold mining would increase. Working around the production possibilities frontier toward more gold and less of commodities would depress the relative price of commodities (remember that the absolute price of gold is fixed), which reduces the demand for money (both the stock of monetary gold and replacement demand for “evaporated” coinage) but increases the demand for commodities. Aggregate income could rise or fall, depending on the relative magnitudes of the changes in the price ratio and the increment to gold production. Remember: with a given production possibilities frontier, we can’t have both more gold and more commodities; we can have more of only one of them. If gold production rises, commodity production – and demand! – must fall, but with a lower relative price of commodities and an unchanged user cost of monetary gold, the only change that can yield a reduction in commodity demand is a fall in aggregate income measured in terms of gold.

But the demand for money also is a function of the relative price of commodities and gold, and the fall of this price ratio will depress the demand for money, but it is not possible that the full amount of the increase in nonmonetary gold demanded could come out of monetary gold, with no change in the quantity of commodities consumed. The only way to squeeze gold from monetary to nonmonetary uses is to change one of the determinants of money demand: aggregate income or the commodity/gold price ratio. The only way for an increase in the demand for nonmonetary gold to affect aggregate income is to change either the relative price of commodities and gold or the quantities of gold and commodities produced. But either of these changes produces a change in the other, so it really doesn’t matter which way we enter this nexus. When all

is said and done (the economy is re-equilibrated with more gold going to nonmonetary uses), the relative price of commodities will definitely be lower *and* aggregate income measured in terms of gold must be lower. If aggregate income were higher than before the change in demand for nonmonetary gold, both determinants of commodity demand would have changed in directions that would raise commodity demand, but we know that if income were to have risen, the price ratio must have fallen to equilibrate money demand, and that would require an increase in gold production and decrease in commodity production to equilibrate the marginal conditions for efficiency in the production side of the economy.

An improvement in gold mining techniques, with a fixed gold price, would initially increase the marginal productivity of resources used in mining relative to those used in commodity production, requiring simultaneously a rise in the relative price of commodities and a shift of resources from commodities to mining to re-equalize marginal production costs. Aggregate income, measured in gold, would most likely rise (only if the reduction in commodity production were large enough to outweigh the increase in the relative commodity price and the increase in gold production would income fall; a country characterized by sharply decreasing returns to inputs into agriculture – most commodities being agricultural in the cases we are interested in – would not have to reduce commodity production by very much to equilibrate the production side of the economy). Both the rise in the relative commodity price and the income increase would increase the stock demand for money, but the effect on commodity demand would be ambiguous, the former depressing commodity demand and the latter elevating it. Some proportion of any increase in aggregate income could be expected to go to the demand for each available good, but in the case of commodities, the income effect would have to offset the price effect for commodity demand to rise. We might well expect to find a greater use of nonmonetary gold, a larger gold money stock, and lower consumption of commodities in a country experiencing a

technical improvement in gold mining – if the country is not able to export the additional gold. The increase in the availability of gold would be unlikely to be neutral in the sense of leaving the real economy unaffected.

9.5.2 *Creation of money by banks*

Private banks have been able to expand their societies' money supplies by issuing notes themselves that serve as currency and by the combined action of accepting demand deposits (checking accounts) and making loans on the basis of those funds. We will not pursue private banknote issue here, but we will show the mechanics of money creation through fractional reserve lending. To understand the process whereby lending creates money, it is useful to begin with a simplified description of a bank's balance sheet in Table 9.1, which lists a banking firm's assets (items due to the bank) and liabilities (sums the bank owes to others).

Reserves would include currency and possibly unminted bullion held by ancient banks. In contemporary accounts they also include the deposits of one bank with another bank, and this practice was widespread in Late Republican and Imperial Rome where loans expanded the money supply based on precious-metal currency to an extent for which evidence has not been found (Harris 2008, 186–188) and seems to have occurred in the fifth and fourth century Athenian banks characterized by Cohen (1992, 2008). Such interlocking deposits among banks are one source of potential instability of individual banks and banking systems, as will become clearer as we proceed with our description of the components of bank balance sheets. Loans are simply the volume of lending made to firms and individuals (and possibly governments).

Both of these components are things that the bank either has or is owed. On the liabilities side, deposits are sums of money the bank has accepted from individuals and firms (and possibly government); the bank holds these sums on behalf of the depositor, but they do not belong to the bank. Deposits are essentially loans from these agents to the bank, whether those lenders receive a positive, zero, or negative interest rate on these deposits. Different types of deposits (defined by the contract characterizing the terms of the deposit) have different maturities, or dates at which the lender can get them back from the bank without incurring a penalty. Demand deposits, commonly called checking accounts today, are “due on demand”: that is, whenever the deposit owner wants to withdraw some or all of his deposit, the bank is obligated to pay the full amount of the request immediately. If the bank does not have sufficient assets on hand to meet this obligation, subject to the banking laws, the bank may have to cease operation and divide up its assets to pay off its creditors. Alternatively, a solvent but temporarily illiquid bank may be able to borrow on emergency terms from some other bank to cover a temporary shortfall in reserves in the face of unexpectedly heavy withdrawals. Equity, the other category of liabilities, is net worth of the firm, held to protect the interests of its creditors in case the bank suffers losses from unpaid loans. The balance-sheet identity is $R + L = D + E$, which says that the value of assets equals the value of liabilities.

Deposits are worth a few more words. If demand deposits can be used to settle debts on the same terms as money through writing the equivalent of checks, demand deposits are effectively part of the narrowly defined money supply. As we noted above, deposits with lesser degrees of liquidity, such as savings and time deposits, may be included in wider definitions of money. In contemporary legal environments, banks are defined by their issuance of demand deposits. Demand deposits have a qualitative difference from other types of longer maturity deposits in that they are an element in the production process that can be called “asset transformation,” which is the transformation of the

Table 9.1 The structure of a bank balance sheet.

<i>Assets</i>	<i>Liabilities</i>
reserves (R)	deposits (D)
loans (L)	equity (E)

maturity structure of assets for the nonbanking sectors of the economy – nonbank production firms and households. We will describe this transformation process in more detail immediately below, but we point out that this kind of transformation is a technical production process just as much as using labor to put seeds into the ground, waiting 6 months and collecting grown-up plants is the production process we call agriculture. The production of demand deposits can increase the liquidity of the economy (under some conditions, it won't have a particularly large effect though) and the money supply. Other types of deposits do not have this property.

Demand deposits are particularly risky to banks since the withdrawal rate of funds from them is only imperfectly predictable. Withdrawal rates depend on the cash needs of depositors, which in turn depend on the temporal pattern of payments by households and businesses. These individual depositors may know quite a bit about their own temporal cash flow patterns, but in general this information is not available to the banks where they hold their deposits, although over some period of time, the bank well may be able to learn a good deal about individual depositors. This nexus of asymmetric information (depositors know more about their cash flow patterns than banks do) regarding withdrawals from demand deposits is one of the principal sources of systemic instability in banking: under the right (unfortuitous) conditions, this situation can produce bank runs, in which depositors fear that there won't be enough reserves in their bank to pay off their deposits in full unless they withdraw them immediately, and of course attempting to do so fulfills the fear. Once this belief becomes current among depositors, it is reasonable for each depositor to attempt to withdraw his deposits before other depositors do, so as not to be the first depositor in line when the vault runs out of cash and the bank closes. This situation, when depositors panic and all try to get their money out of bank deposits at the same time, is called a bank run (or a run on a bank). Bank runs are costly in both financial and real terms. To the extent that

banks try to sell off their assets prior to maturity (call loans, sell off other loans or securities), they get less than their value; to the extent that a lot of banks put their units of particular assets on the market at the same time, they drive down the sale price (demand for them remains about constant, but the supply curve shifts way out to the right). These asset price changes impose financial losses. Real losses are inflicted on depositors who were planning real production activities on the basis of their anticipations about having cash on hand. Some of these production activities will be cancelled; some materials will spoil or otherwise become unusable, and so on. These are the real losses imposed by a bank run.¹¹

Returning to the balance sheet, a bank takes in deposits on its liability side (effectively taking out a loan from depositors) and has to do something with the asset side of its balance sheet to keep the identity in balance. The deposits go either into the bank's reserves or its loans, or a combination of the two. The only way it will make any profit (or even the normal rate of return on the bank's inputs and capital) is to loan some or all of the funds taken in through deposits to other firms or individuals at an interest rate higher than the interest rate it pays on deposits (which may in fact be negative, representing a service charge). Depending on the relationship between reserves and loans on the asset side of the balance sheet, the deposit-lending process, as it works its way through other banks in the economy, will increase the demand-deposit component of the money supply. Using the contemporary convention of a reserve requirement, we proceed to show how this mechanism works. In the subsequent subsection we will discuss the behavioral underpinnings of what is presented rather mechanically here.

We will work through two cases of money creation by banks. In the first case, we will assume that bank money is the only money – that is, there is no currency – and that the only type of deposit is the demand deposit. In the second example, we introduce currency holdings and show how the demand for currency restricts the money creation

Table 9.2 Bank money creation 1: the initial balance sheet of Bank 1.

<i>Assets</i>	<i>Liabilities</i>
reserves 2 talents	deposits 10 talents
loans 8 talents	

of banks. When we have finished working our way through a few tables of numbers describing the progress of loans and deposits through a banking system, we will furnish some coefficients known as multipliers that relate changes in each of the balance sheet entries to some initial change in one of them. In the hypothetical balance sheet for Bank 1, in Table 9.2, we exclude the equity on the liability side to highlight the interaction among deposits, loans, and reserves, the fulcrum of the money-expansion process.

To simplify how we think about the relationship between deposits and reserves, we assume that the banking system has the contemporary convention of a minimum reserve requirement: all banks must hold reserves equal to at least 20% of the value of their demand deposits to ensure that they do not get caught short by unanticipated withdrawals of deposits. In reality, even with a legal minimum reserve requirement, many banks would want to hold more than the minimum reserves required, or “excess reserves.” In this example, we assume that no bank wants to hold excess reserves. To kick off the monetary change, let’s suppose that the government had borrowed 1 talent from Bank 1 and repays it in silver. This reduces Bank 1’s loans by 1 talent and increases its reserves by 1 talent, keeping the asset side of the balance sheet constant, but leaving the bank with 1 talent more in reserves than it has to maintain, as shown in Table 9.3.

Table 9.3 Bank money creation 2: government pays off a loan to Bank 1.

<i>Assets</i>	<i>Liabilities</i>
reserves 3 talents	deposits 10 talents
loans 7 talents	

Table 9.4 Bank money creation 3: Bank 1 loans out its excess reserves from the government loan payment.

<i>Assets</i>	<i>Liabilities</i>
reserves 3 talents	deposits 11 talents
loans 8 talents	

Immediately, Bank 1 has 1 talent of excess reserves. Since we have assumed that it desires no excess reserves, it loans the extra talent out, adding 1 talent to loans and adding one talent to deposits on the liability side of the balance sheet. To effect the loan, it adds one talent to the demand deposit account of one (or more) of its deposit holders: see Table 9.4.

The borrower who had the loan of 1 talent placed in her demand deposit account spends the borrowed money by writing a check against her deposit. Suppose there are lots of banks – so many that the chance of the person to whom this borrower wrote the check using the same bank is just about zero: the check gets deposited at another bank – Bank 2 – and Bank 2 sends the check to Bank 1 to get the money for its own depositor. Bank 1 immediately loses the full amount of the extra 1 talent it had in excess reserves, but the amount of its deposits also falls by the same amount: see Table 9.5.

Now we turn to Bank 2 and look only at the changes in its balance sheet. When the fellow who sold a load of lumber to the woman who borrowed the talent from Bank 1 deposits the check in his bank – Bank 2 – Bank 2’s deposits go up by 1 talent; Bank 2 sends the check over to Bank 1 for settlement, and as soon as the cash equal to the

Table 9.5 Bank money creation 4: Bank 1’s depositor buys goods with his loan, lowering Bank 1’s deposits, and Bank 2 sends the check to Bank 1 for its settlement, reducing reserves.

<i>Assets</i>	<i>Liabilities</i>
reserves 2 talents	deposits 10 talents
loans 8 talents	

Table 9.6 Bank money creation 5: changes in Bank 2's balance sheet when it receives the check for the amount loaned out by Bank 1 and clears the check.

<i>Assets</i>	<i>Liabilities</i>
reserves +1 talent	deposits +1 talent
loans no change	

Table 9.7 Bank money creation 6: Bank 2 loans out its new excess reserves.

<i>Assets</i>	<i>Liabilities</i>
reserves +0.20 talent	deposits +1 talent
loans +0.80 talent	

amount of the check reaches Bank 2, its reserves go up by 1 talent as well, so the two sides of its balance sheet remain in balance (Table 9.6).

At this point, Bank 1 is back to its legal minimum reserve requirement, but Bank 2 now has excess reserves (before this check came in, Bank 2 was just meeting its reserve requirement). Bank 2's demand deposits have increased by 1 talent, so its reserve requirements have increased by 0.20 talent. Thus, Bank 2 finds itself with 0.80 talent of excess reserves, which it will want to loan out. The composition of the liability side of Bank 2's balance sheet changes accordingly, as shown in Table 9.7.

Bank 2 puts its loan of 0.80 talent into a demand deposit for its borrower. This borrower also spends the loan immediately so he can put the funds to work – no point in paying interest for nothing. He buys a load of bricks from a brickmaker who banks at Bank 3. The brickmaker puts the check for 0.80 talent in his account at Bank 3, which now has both sides of its balance sheet increase by that amount (it has already sent the check back to Bank 2 for the cash promised by the check): see Table 9.8.

Bank 3, which also was holding its minimum reserves, now finds itself holding excess reserves: its deposits increased by 0.80 talent, which requires the addition of $0.80 \times 0.2 = 0.16$ talent in additional reserves while it finds itself

Table 9.8 Bank money creation 7: Bank 2's loan arrives at Bank 3.

<i>Assets</i>	<i>Liabilities</i>
reserves +0.80 talent	deposits +0.80 talent
loans no change	

Table 9.9 Bank money creation 8: Bank 3 loans out its new excess reserves.

<i>Assets</i>	<i>Liabilities</i>
reserves +0.16 talent	deposits +0.80 talent
loans +0.64	

with 0.8 talent in reserves. It immediately lends out its excess reserves of 0.64 talent, and the asset side of its balance sheet changes as shown in Table 9.9.

This process of receiving new deposits and reserves and lending out the excess reserves continues until all the new reserves have been fully absorbed into required reserves – until there are no more excess reserves. Each step in the expansion is equal to a factor of one minus the reserve requirement times the volume in the previous step. If we add all these steps up, we have a geometric progression which, by the time all the excess reserves have been absorbed, yields a total volume of new deposits equal to the initial change in the balance sheet that set off the expansion times the inverse of (“one over”) the required reserve ratio: $\Delta D = \Delta B/r$, where Δ means “change in,” r is the fractional reserve requirement expressed as a decimal fraction (0.20 in our example), and B is the monetary base, a term that we have not explained yet but will immediately. With a reserve ratio of 0.20, the increase in the money supply will be five times the change in the monetary base. The 1 talent increase in reserves at Bank 1 will have been expanded into a 5 talent increase in the money supply through the loan-deposit creation process – assuming that these deposits are indeed demand deposits in the full meaning of the term. If the only way banks can make

loans is to issue coins to borrowers, there is effectively a 100% reserve system, and consequently no monetary expansiveness to the banking system.

The monetary base is the money against which loans may be made in a banking system: in this case, currency denominated in talents and located in the reserves of the banking system. Another term you will encounter for the monetary base is “high-powered money,” which is more descriptive of the capacity of this component of the monetary system to expand the total money supply in multiples of changes in its own size. In the absence of a treasury (government’s presence) or a central bank in a society’s monetary system, this monetary base or high-powered money would simply be the quantity of metal currency and bullion in the reserves (vaults, holes in the ground, and so forth) of private banks. If the government operates a treasury, and if the treasury uses its precious metal holdings as currency, the economy’s high-powered money will be the sum of the treasury’s precious metal holdings and the currency (precious metal, coin, dumps, and so forth) held by private banks.

Now, consider what happens to this process if banks either want, or are required, to hold 100% reserves against their demand deposits. We can do this either of two ways: (i) go back to Table 9.2 and see what happens when Bank 1 gets its infusion of new reserves, see how much it can loan out, and trace the loans and new deposits through the banking system, or (ii) substitute $r = 1.0$ into the formula $\Delta D = \Delta B/r$ to get $\Delta D = \Delta B$. The monetary base will have absolutely no multiplicative expansive capacity.

Let’s proceed to put currency back in our monetary and banking system, since we know that, by the time monetary systems emerged in the ancient Mediterranean and Aegean countries, they were generally metal coins. In our first case, in which the monetary base was composed entirely of bank reserves, we could define the monetary base as $B = R$, where R stands for bank reserves. Now we let people outside of banks hold metal currency too, and the monetary base becomes defined as $B = R + C$, where C is currency in circulation with the nonbank public. Recalling that the

volume of bank reserves is equal to the volume of demand deposits times the reserve requirement on those deposits, or $R = rD$, we can substitute into our new definition of the monetary base to get $B = rD + C$. We can rearrange this expression to find out the volume of demand deposits in terms of the monetary base: $D = (B - C)/r$; or in terms of changes, $\Delta D = (\Delta B - \Delta C)/r$. When the public wants to hold currency as well as demand deposits, the ratio of deposits to the monetary base shrinks. If the monetary base increases but public currency holdings remain unchanged, $\Delta C = 0$, and we have the same expansiveness of deposits relative to the change in the monetary base: $\Delta D = \Delta B/r$. But it’s not unreasonable to suspect that the public might want to maintain a roughly constant ratio between its holdings of demand deposits and currency, say a 5-to-1 ratio of deposits to currency. If we express the currency / deposit ratio as c , we have $C = cD$, and $\Delta C = c\Delta D$. Putting this expression back into the formula for bank reserves, we get $\Delta B = r\Delta D + \Delta C = r\Delta D + c\Delta D = (r + c)\Delta D$.

Expressing the change in deposits as an expansion of base money, we now get $\Delta D = \Delta B/(r + c)$. As $r + c$ is larger than just r , the expansiveness of base money into deposits is smaller than when the public didn’t hold currency in addition to deposits. Instead of a fivefold increase in deposits following our original one-talent infusion of reserves into Bank 1, we would get a two-and-a-half-fold increase with this ratio of currency holding. This drain of currency from the banking system reduces the volume of reserves available to the banking system on which it could make loans and expand deposits.

Before departing this rather mechanical approach to loan expansion by a banking system, let’s put equity back into the liability side of the balance sheet for a moment. Typically banks will want to maintain some ratio of equity to loans to protect their creditors against unexpected loan losses. Just as we expressed the (assumed mandatory) ratio of reserves to deposits as $R = rD$, we can express the desired ratio of equity to loans as $E = eL$. Using the balance sheet identity, $R + L = D + E$, we can express the changes in

the elements of an individual bank's balance sheet in terms of inflows of deposits it receives. Again, assuming no redeposit of checks at the issuing bank, we have an equity multiplier of $E = [e(1-r)/1-e]I$, where I is the inflow of deposits, a deposit multiplier of $D = (1-e/1-e)I$, a credit or loan multiplier of $(1-r/1-e)I$; and a reserve multiplier of $R = rI$. Looking at the loan multiplier and comparing it to the expansiveness we found in our simple formulation above, allowing for banks' caution in maintaining a desired equity-loan ratio reduces the expansiveness of any given deposit inflow, and hence of any infusion of high-powered money into the banking system.

9.5.3 *The banking firm*¹²

The lending and reserve behavior described mechanically in the previous subsection will be conducted, in reality, in a fashion to maximize the profit of the banking firm. How much lending it does out of excess reserves will be a profit-maximizing choice. Before showing the structure of a profit-maximizing bank's operations, we discuss the technical production that a bank accomplishes. We can divide the agents in an economy, for the purposes of thinking about banking, into several "sectors": the household sector, the nonbank production sector (construction firms, shipping firms – both sea and land – potteries, leather-goods firms, merchant houses, and so on), and financial intermediaries – firms that seek to connect borrowers and lenders with specialized needs and capacities. Financial intermediaries on a large, industrialized scale, are a relatively recent development, but it is likely that such specialized financial activity long has been conducted on a more sporadic basis and smaller scale, possibly scattered among other activities within a single (family) firm, likely not especially well served by contract law in antiquity. These activities may transform assets of one degree of liquidity or maturity into assets of another type, although they may simply act as brokers, middlemen or dealers in similar types of assets and liabilities without transforming their degree of moneyiness.

Nonbank financial intermediaries, the subject of the following sub-section, typically increase the liquidity of the economy but reduce its money supply, since they hold some monetary assets as reserves while even their most liquid assets, such as savings deposits, are not money, at least in the narrow definition.

Banks are one type of financial intermediary firm. They typically convert less liquid assets into more liquid ones, and assets with less predictable values into ones with more predictable values. Using the standard factors of production available to other businesses – labor and capital – they solicit short-term deposits to turn into longer term commercial loans and mortgages, increasing the liquidity of households and production firms and usually increasing the money supply, although their ability to do that depends on a number of factors including legal ones. Banks supply money to the rest of the economy if their monetary liabilities – principally demand deposits – exceed their monetary assets such as currency reserves. Banks could, conceivably, reduce the money supply of the economy if their balance sheet structures have the reverse composition.

In their asset transformation activities, banks encounter three principal types of risk – uncertainty about future interest rates, important because of the shorter maturity of their liabilities than their assets; the uncertain maturity of their deposits; and the possibility of bad loans. Banks essentially borrow short (their deposits) to lend long (their loans), thus creating greater liquidity for their clientele and leaving themselves with less liquid balance sheets. In doing so, they must finance longer maturity loans with short-term loans whose future cost is unknown at the time of incurring the longer term loan. If short-term interest rates are about the same as long-term rates (a relatively flat term structure of interest rates, or "yield structure") and there is considerable uncertainty about the future, or if bankers are highly risk averse, the maturity structure of banks' assets and liabilities will be closely matched – that is, their loans won't be of much longer term than their deposits, and their asset transformation will be slight. On the

other hand, if circumstances in the economy put a premium on short-term funds (giving higher short-term interest rates than long-term rates), and the uncertainty about future rates is not particularly great, or if bankers are more risk neutral (or oblivious), banks will have a more marked tendency to borrow short and lend long, increasing the economy's liquidity.

The uncertain maturity of deposits puts banks at risk of running out of reserves if they face unexpectedly large withdrawals. The cost of keeping reserves is the foregone interest cost on loans, but the cost of running out of reserves can be high – penalty rates on emergency loans or even going out of business for either illiquidity or insolvency. The bank will strive to balance the potential marginal loan revenues against the expected marginal cost of running out of reserves. This balancing act creates a demand for reserves by individual banks, even in the absence of legal minimum reserve requirements, and limits the asset transformation banks are willing to accomplish.

To the extent that the probabilities of default losses are known from experience, they can be treated like any other cost – labor and capital equipment. The interest rate on loans will be raised accordingly. To the extent that they are uncertain, this risk forces banks to maintain equity capital which could otherwise earn a return. They will hold equity so as to balance the cost of this capital against their need to protect their creditors against loan losses and possible bank failure. This is the source of the coefficient e in the last paragraph of the preceding subsection. These three sources of uncertainty facing banks are constraints on a bank's profit maximization problem, to which we now turn.

Consider the structure of the banking firm's profit maximization problem. The source of income for the bank is what it makes off its loans, less what it loses from defaulted loans: $(i_L - \beta)L$, or the difference between the loan rate and the percentage of defaults, times the loan volume. The bank's ordinary factor inputs are labor N and capital equipment K , for which it pays wage w and capital rental r . It also must pay something for its deposits, either a positive interest rate or

a negative service charge. It also pays a normal rate of return for the equity on the liability side of the balance sheet. Finally, the bank may occasionally find itself short of reserves but be able to borrow at some emergency, or penalty, rate i_p to cover a short-term liquidity problem of expected size ρ . The bank's expected profit then is $E(\Pi) = (i_L - \beta)L - wN - rK - i_D D - eE - i_p \rho$. The firm adjusts both its balance sheet, containing L , R (implicitly contained in the relationship represented by ρ), D , and E , and its factor inputs, N and K .

By manipulating these balance sheet entries and factor inputs, the bank maximizes this expression for expected profit subject to three constraints: Its production function describes how it combines capital and labor to assemble deposits and transform them into loans: expressed as an implicit function, $f(N, K, L, D) = 0$, in which N and K are conventional factor inputs and L and D are outputs. The production function does not involve risk, but the other two constraints characterize how the bank adjusts its balance sheet items to take account of the risks of deposit withdrawals exceeding reserves and of the bankruptcy probability exceeding a given magnitude. The withdrawal risk constraint appeals to the relationship captured by ρ and can be expressed as $\rho = \rho(D, R)$ (read "the expected value of reserve shortfalls ρ is a function, called ρ , of the volumes of deposits and reserves"). Larger D increases ρ , and larger R decreases ρ ; in words, a larger volume of deposits, holding reserves constant, increases the probability that the bank will sooner or later run short of reserves; and holding a larger volume of reserves, for a given loan volume, reduces the probability of a reserve shortfall. The other probabilistic (stochastic) constraint lets the bank set the expected value of bankruptcy loss caused by loan defaults that it is willing to live with: $\bar{\omega} = \omega(L, E)$ (read "omega-bar, the 'bar' denoting a fixed value of omega, is a function omega of loan volume and the volume of equity in the bank"). The expected size of bankruptcy loss rises with larger L , equity held constant, and decreases with larger equity, loan volume held constant. The only way to drive bankruptcy risk to zero is

to hold reserves equal in volume to loans; that would be very costly in terms of foregone interest earnings.

Recall from earlier chapters how we formed Lagrangean expressions and varied the magnitude of the choice variables to find the maximum (or minimum) value of the objective function. In this case the choice variables are N , K , L , D , R , and E , and the profit formulation is the objective function to be maximized. The first-order conditions are the relationships among these variables that identify their profit maximizing values. From the first-order conditions for labor and capital, we get the result that those two inputs should be used in a combination that equalizes the marginal costs of loans in terms of both factors. In other words, the marginal cost of a value of loan should be the same in terms of either conventional factor of production. This is the banking equivalent of the ratios of marginal productivities and marginal costs being equated for all inputs in an agricultural or manufacturing production process.

Two other first-order relationships relate the marginal revenue from loans to their total marginal cost and the marginal benefit of holding reserves to their marginal cost. We saw the marginal benefit of loans in the expected profit formulation as $i_L - \beta$ and the marginal cost of not having sufficient reserves, in the form of the penalty rate i_p . The marginal cost of another unit of loan has three components – the marginal factor cost of the loan (the labor and capital cost of researching, packaging, and placing it); the marginal equity cost (the amount by which equity needs to be raised to keep the expected bankruptcy value at the desired level); and the expected marginal penalty cost (any increase in loans not financed by new equity comes out of reserves, which increases withdrawal risk, or what we have called insufficient reserves risk).

The condition for the optimal amount of deposits, for a fixed level of loans equates the marginal benefit and marginal cost of reserve holding. That is, if we equate the marginal benefit of holding reserves to their marginal cost, we will have found the profit-maximizing volume of loans. The marginal benefit from holding an

additional unit of reserves is the reduction in the expected penalty cost of a reserve shortfall. The marginal cost of providing this extra unit of reserves consists of the factor cost, the interest cost of assembling another unit of deposits and the expected additional penalty cost of a reserve shortfall caused by the additional unit of loan.

These three relationships yield demands for deposits, loans, reserves, equity, capital goods, and labor as functions of interest rates and factor costs. If i_L rises, the volume of loans and, consequently, equity requirements rise; the demand for deposits rises because it becomes more profitable to transform deposits into loans with a higher loan rate. The effect on reserves is ambiguous. The substitution of loans for reserves reduces reserves (recall that there is no legal minimum reserve requirement in this analysis, only a profit-maximizing level of reserves), while the overall scale of bank operations increases reserves. The average reserve rate falls since the higher loan rate makes it profitable to accept a higher expected reserve shortfall, as the statistical law of large numbers makes it possible to achieve the same expected shortfall with a lower reserve/deposit ratio.

A higher deposit interest rate causes the volume of deposits to contract – remember that the deposit rate is an outflow for the bank. Loans and equity contract correspondingly. Reserves decline but less than proportionately to deposits because of the reverse effect of the law of large numbers. The reserve ratio rises. A higher penalty rate increases reserve volumes and the reserve/deposit ratio. The effect on deposits is ambiguous. Loans and equity contract because it becomes more profitable to hold funds than to lend them.

This effect of the law of large numbers on profit-maximizing reserve ratios highlights an important source of scale economies in banking: the probability of bankruptcy will fall with bank size. Consequently the size of banks may be limited by their markets, including transportation costs. Such scale economies appear to be larger for demand deposits and commercial loans than for mortgages and time deposits, and different size banks accordingly will have different balance-sheet compositions. Smaller banks will

concentrate on lower-risk assets and liabilities and on smaller sizes of loans and deposits. They will be somewhat more labor intensive, using the labor to deal with smaller accounts, both of deposits and loans.

Transplanting this model of an individual banking firm to an entire economy with many banking firms as well as households and nonbank businesses, we can address what the existence of a banking system does for an economy, in terms of affecting interest rates and prices and the aggregate money supply. Since this analysis involves the behavior of households and nonbank production firms as well as all banks in the economy, it is considerably more complicated (bigger – more behavioral and accounting relationships to be embodied in equations and variables). Consequently we do not show the entire model here, but only the portions of it relating to the banking sector.¹³ Continuing with the institutional case of no legal reserve requirement and ignoring loan defaults to simplify the problem a bit, we can reformulate the bank's optimization problem as $\text{Max } E(\Pi) = i_L - i_D - \delta R - wN_b - rK_b - i_p p_x \rho$, subject to $L + R = D$ and $\bar{p} = (1/p_x)W(D, R, N, K)$. We have ignored equity in this problem; δ is the rate of loss of gold currency due to abrasion, loss, and so forth; the subscript b on N and K indicates that these variables represent the employment of labor and capital equipment in the banking sector, with the nonbank business sector employing the remainder available in the economy. The expected reserve shortfall, represented by ρ , is measured in terms of commodities, at the price of commodities, p_x . The function $W(\blacksquare)$ in the reserve shortfall constraint is the risk of unexpected withdrawals causing a bank to run out of reserves; letting $\Delta W/\Delta D$ and so on be represented by the notation W_D , and so forth, the relationships in that function are $W_D > 0$, $W_R < 0$, $W_N < 0$, and $W_K < 0$, which means the following: increasing the volume of deposits, everything else held constant, increases withdrawal risk; increasing the size of reserves, all other inputs and outputs held constant, reduces withdrawal risk; and using more labor and capital, for instance to package deposits and loans in smaller, more bite-sized units, reduces withdrawal risk. The bar over p

in the second constraint indicates that banks are picking a specific level of withdrawal risk to live with.

Households, for which we do not show the specific behavioral relationships, in maximizing their utilities, arrange their own asset holdings between currency and deposit money so as to equalize the marginal utility they receive from each asset stock. In the meantime, competition among banks will induce them to keep a reserve ratio on their marginal deposits equal to households' marginal rates of substitution between currency and deposits, which also happens to be the ratio of (i) the marginal benefits of deposits in terms of the loans they permit to (ii) the marginal cost of reserves held on the last unit of deposits accepted from households or businesses, in terms of foregone interest and the holding cost of gold currency. In symbolic terms, this relationship is $\text{MFR} = (i_L - i_D)/(i_L + \delta)$. If we take the numerator of the right-hand side of this relationship over to the left-hand side and multiply it by the marginal reserve ratio (MRR), we get the unit cost, in terms of foregone loan and user or holding cost of gold currency, of the last unit of deposits held by the bank – a price times a quantity. This demonstrates the connections between households, which go about their business with only a few points of contact with banks, and the banking system.

Across the entire economy, the issuance of money by banks reduces the demand for currency. Additionally, the composition of output by private, nonbank production firms shifts from gold to commodities. As we saw in the case of technical progress in gold mining in section 9.5.1, this shift in the allocation of productive resources is accomplished by a rise in the price of commodities. While factors of production (labor and capital) move from gold mining to commodity production, there is a similar shift of labor and capital from commodity production to banking, and total output in the production sector (gold plus commodities) falls relative to the production of the banking sector. The final effect on commodity output is ambiguous (that is, we can't say in general whether it will rise or fall; that depends on the relative strengths of various effects): the

composition effect increases it while the overall reduction in output reduces it. The increase in commodity prices depresses the real value of the gold stock. Without banks, deposits were nonexistent, which simply says that the interest rate on deposits was so low that nobody was willing to put any wealth into that type of asset. With the emergence of banks, the positive volume of deposits (up from a zero volume) implies that the interest rate on deposits rises. Conversely with loans, and the rate on loans with the existence of banks is lower than without them, and we see a positive volume of loans in the banking equilibrium. We cannot say that, in general, the existence of the financial intermediation provided by banks will decrease interest rates: it lowers some and raises others.

As long as the bank deposits are convertible into gold at a fixed price, the absence of a minimum reserve requirement for the banking system does not cause the price level to go off to infinity. As in section 9.5.1, the price level is determined by cost conditions in gold mining. Although the commodity price level, in terms of gold, will have the formal relationship of being higher, the higher the ratio of gold to goods, the quantity-theoretic appearance of the price level being positively related to the quantity of money, is derivative rather than causal.

9.5.4 *Financial intermediation*

Financial intermediation puts together borrowers and lenders so the borrowers can gain access to resources in excess of their current supplies and the lenders can find profitable investments for resources they do not want to consume or hold in the present period.¹⁴ The bank may be the first type of institution that comes to mind as a financial intermediary, and banks do provide financial intermediation, but they also create money as discussed in the previous subsection, while this subsection focuses on non-monetary (nonbank) financial intermediation.¹⁵ Intermediation goes beyond trade in loanable funds via face-to-face negotiations between ultimate borrowers and ultimate lenders, as well as self-financing, generally by the wealthy. Both

types of nonintermediated financing restrict the opportunities investors find and in general retard the economy-wide efficiency of investment and capital formation. This section will use a number of anachronisms simply because it makes the exposition easier. Some scholars believe that true financial intermediation in the ancient world was quite limited, and surely it was in many places and times, even if it was possibly not quite as rare as has sometimes been thought.¹⁶ My approach in this section is to offer the reader information on the structure of financial intermediation on the possibility that specialists in ancient finances may notice some structural similarities to situations that the evidence from antiquity has yielded in one form or another. I will attempt to connect the structures of financial intermediation to ancient examples where I can, but where such connection is not possible, because of either the facts of antiquity or my knowledge of them, the anachronisms will have to suffice.

The introduction of some terminology will help the exposition. First, nonfinancial spending units: these are simply entities that may spend resources. They may be households looking for a mortgage, a consumption loan for a wedding ceremony, a production loan to buy a new draft animal, or to earn some return on resources they do not plan to consume soon rather than sticking them in the ground. They may be businesses – small or large – looking to either loan resources for which they have no immediate spending plans or to borrow resources to replace some equipment or expand operation. Their principal business is not the conduct of financial operations. Intermediation brings together deficits and surpluses among spending units. Second, primary security: this is an asset issued by a nonfinancial spending unit. In contemporary industrialized countries, these could be bonds (loans) and stock (ownership) issued by firms; in antiquity, it could be a cart rented out for a season by a well-to-do farm household; it could be a mortgage taken out by a household for either a house or land. Third, indirect securities: these are obligations of financial intermediaries which have purchased primary securities from nonfinancial spending units. Examples include

currency, demand and savings deposits at banks, shares or deposits at specialized institutions such as land banks and consumer credit unions (definitely an anachronism), insurance policies, and other such claims.

There are two principal types of financial techniques: distributive techniques and intermediary techniques. Distributive techniques increase the efficiency of markets on which ultimate borrowers sell and ultimate lenders buy primary securities. Intermediary techniques bring financial institutions into competition with one another for the purchase of primary securities and substitute indirect financial assets for primary securities in the portfolio holdings of ultimate lenders. These techniques raise the level of saving and investment in an economy and allocate the economy's scarce savings more efficiently than either direct loanable funds trading or self-funding. Distributive techniques include the wide provision of information to borrowers about the asset preferences of lenders and to lenders about the securities issued by borrowers. Technological advances in distributive techniques postdating antiquity include the establishment of brokerage facilities and the provision of facilities for future as well as spot deliveries (for example, the purchase today of next year's wheat crop, to be delivered after next year's harvest, the purchase of goods from a trading expedition that is not expected to return for a year or more). These techniques provide greater diversification of debt or financial assets and access to a wider range of borrowing and lending options. Intermediary techniques include the repackaging of heterogeneous obligations issued by small borrowers into homogeneous, standardized securities that are more widely marketable. While this particular operation may seem very clearly to be an anachronism, it may be useful for scholars of antiquity to recognize activities the ancients were not able to perform, as well as to understand how the operations they were able to conduct operated, facilitating a clearer recognition of ancient limitations as well as capacities.

Financial intermediation includes the monetary system as well as nonmonetary intermediation firms. Financial intermediation through the

monetary system can work more smoothly and extensively when a monetary authority can buy private or government securities, thereby issuing money, as well as conduct the reverse operation, contracting the money supply. However, structurally, ancient governments got their money into circulation by making purchases, sometimes of assets, but probably more frequently of consumable goods and services. Financial intermediation through the monetary system transfers credits between spending units. The rest of this section will focus principally on non-monetary financial intermediation, that is, the operations of firms that issue securities other than money.

Nonmonetary financial intermediaries (NMFI) purchase primary securities from ultimate borrowers and issue indirect debt in the form of nonmonetary claims on themselves to be held in the portfolios of ultimate lenders. The principal asset in a NMFI's portfolio is primary securities, but they also can hold indirect securities of other intermediaries as well as tangible assets of their own. Contemporary NMFIs include both private and public institutions. Among the private types of NMFI firm are savings and loans, mutual savings banks, life insurance companies, credit unions, and other specialized agencies. Among government types are agricultural land banks, which actually could be formed privately. Focusing on private institutions henceforth, most NMFIs purchase a relatively narrow band of primary assets – for example, mortgages, corporate equities or bonds, government debt, and so on – although some purchase a wider range. Most NMFIs also issue a correspondingly narrow range of indirect debt, although some, such as insurance companies, issue a much wider array. Some, such as sales finance companies, frequently rely on other intermediaries to buy their debt.

The product of intermediation is the indirect financial asset constructed from the underlying primary security. The reward from intermediation derives from the difference between the rate of return on primary securities the intermediaries hold and the interest or dividend rate they pay on their indirect debt. The indirect securities

provide different service flows, such as access to mortgage funds or consumer loans, the security against misfortune provided by insurance companies, opportunities for diversification provided by mutual funds (surely another anachronism, although consideration of the security packaging a mutual fund provides might offer some useful analogies). The similarities among indirect securities include easily identifiable redemption values, low investment costs, and divisibility into convenient units.

Financial intermediaries exploit scale economies in borrowing and lending. This requires fairly high specialization in both their assets and liabilities. On the lending side, intermediaries can place and manage investments in primary securities at costs below that most individual lenders would incur. The size of the typical intermediary's portfolio allows the risk reduction provided by diversification, and intermediaries can schedule the maturities of their debts to reduce the chance of experiencing a liquidity crisis. On the borrowing side, an intermediary with a large number of depositors can rely on some predictability in the timing of claims for repayment, allowing it to get by with a relatively illiquid portfolio that yields a higher return. These advantages are distributed among the intermediary's debtors in the form of more favorable loan terms than they could obtain otherwise and to its creditors in the form of higher interest payments than they would otherwise receive.

Primary security issues can operate through three channels, one direct and involving no intermediation, two indirect and involving intermediation. First, they can be sold directly to other nonfinancial spending units, in direct finance. Second, they can be sold to the monetary system, the ultimate lenders acquiring money rather than primary securities. Third, they can be sold to NMFIs, in which operations the ultimate lenders acquire nonmonetary indirect assets instead of primary securities or money.

Spending units' demands for nonmonetary indirect assets depend on their income, their holdings of financial assets, the interest rates on primary securities, and the deposit rates paid on money and by NMFIs on their indirect debts.

An increase in income has offsetting effects on the demand for nonmonetary indirect assets; it increases the demand for money, which reduces the demand for all other financial assets, while the demand for certain types of nonmonetary indirect assets may parallel changes in income. Higher interest rates on primary securities reduce the demand for nonmonetary indirect assets because they make the primary securities cheaper, while higher deposit rates paid by NMFIs increase the demand for their indirect securities.

Nonmonetary financial intermediaries' willingness to supply claims on themselves depends on the interest rates on primary securities, the deposit rates they pay on the indirect debt they sell, their expenses, which vary with the amount of business they transact, and on the types of primary assets available. Increases in the interest rates on primary securities shift out NMFIs' supply curve of nonmonetary indirect assets because the prices of those securities will be lower. If they have to pay a higher deposit rate on their own liabilities their supply curve will shift to the left. Higher variable costs will also shift their supply curve to the left, as will a composition of primary assets that is unfavorable to their line of activity.

In the short run, at a given interest rate, an increase in the demand for nonmonetary indirect assets will lower the deposit rate NMFIs have to offer, as shown in Figure 9.3. In that figure,

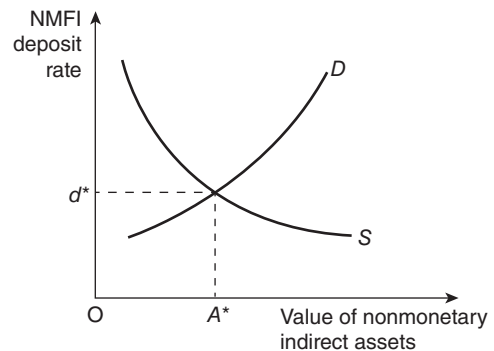


Figure 9.3 Supply of and demand for nonmonetary financial assets.

the slopes of the demand and supply curves are reversed from their usual slopes, because higher deposit rates make these assets more attractive to purchasers – hence the upward sloping demand curve – while higher deposit rates reduce the willingness of NMFIs to offer more debt.

Financial intermediation can increase the supply of loanable funds. When spending units increase their demands for nonmonetary indirect assets, NMFIs must increase their demands for primary assets on which the indirect assets are based. However, an increase in spending units' demands for nonmonetary indirect assets reduces their demand for money holdings, and if the reduction in their money demand exceeds the increase in their demand for the indirect assets, the interest rate will have to fall to equilibrate the money market, otherwise the unchanged stock of money would exceed its demand. This is equivalent to a growth in the NMFI market leading to an increase in the supply of loanable funds, because the interest rate has fallen. Nonmonetary financial intermediaries can further affect spending units' demand for money by altering the composition of primary securities the spending units hold. For example, if NMFIs sold government securities (debt) to spending units and purchased mortgages from those spending units, this switching operation would remove illiquid securities and replace them with more liquid ones (another anachronism, but possibly a useful reference case). Consequently, the spending units could afford to hold smaller money balances. A switching operation could work in the other direction as well, however, with spending units becoming less liquid and thus increasing their money balances to keep their overall liquidity at the desired level.

9.5.5 *Exogeneity / endogeneity of money supply and foreign exchange*

Does anybody control the money supply? Another way to ask this question is, "Is anybody in charge?" "Does anybody *think* he is?" The answers to these questions depend on the institutional setting. If a government operates a treasury, a mint, or both, the answer is, "It depends." With

a metallic standard and the rule of free coinage, which our analysis of this section has used, the answer is, "No." If the government restricted the quantity of metals minted, it could certainly restrict the quantity of coinage, but in doing so it would give up the seigniorage¹⁷ it could make on the minting, and it might find it nearly impossible to restrict the parallel circulation of unminted quantities of metals (dumps, nuggets, "gold dust," and so forth) if a larger quantity of money was demanded than it minted.

Monetary authorities have closer control over fiat money supplies than over commodity money supplies, but even then, once international connections are considered, the control may be limited, if not downright illusive. We introduce a bit of terminology. An exchange rate is the price of one currency in terms of another. If we considered the exchange rates between different precious metal coins, it would be nothing more than the relative metal contents of a coin of each issuer's (country's) unit of account – for example, the gold weight of an Egyptian deben relative to the silver weight of a Phoenician shekel. Translating between the different metals introduces the matter of the relative prices of the metals, but once that is accounted for, the exchange rate is simply a matter of relative weights. A fixed exchange rate is an agreement between countries to fix the values of their currencies in terms of one another's. A flexible, or floating, exchange rate is one that the respective issuers allow to be determined by the market for the two currencies. Under a fixed-rate regime, money flows automatically between (among) countries in accordance with local demands. States have no control over the money supplies within their borders, although they can mint as much as they want – it will just flow out of the country; or "needed" coin will flow in if a country mints too little relative to demand. The local authorities also have no control over their price levels beyond setting the price of gold in terms of the local unit of account. This all gets a bit more complicated when a monetary authority (or private banks) issues credit money, even if the countries involved are on a gold standard. Local credit money issues (call

them “notes” for simplicity) in excess of demand at the equilibrium price level will simply flow back to the issuers for redemption in gold (or whatever metal is the standard), and the gold will be shipped overseas until the multinational equilibrium distribution of gold is reattained. According to the rules of a fixed rate system, monetary authorities are obligated to buy or sell their currency in quantities required to maintain its exchange rate within a small band around par (the agreed upon exchange rate). When a monetary authority has let its money supply either grow or shrink to such an extent that it cannot make the required purchases or sales, it must devalue (lower the price of) or revalue (raise the price of) its currency.

In a fixed-rate system, the term “balance of payments” has an important meaning: the surplus or deficit of exports over imports, in the absence of financial capital flows, is equal and opposite in sign to the value of gold flows used to finance the deficit of whichever country or countries is, in effect, borrowing (spending more than it produces). If a country has a surplus of exports over imports, it is in effect saving – it’s producing more than it consumes. If it has a deficit of imports over exports, it’s borrowing. This mechanism will operate whether a monetary system uses notes in addition to metallic currency or uses only coins.

In a flexible-rate international monetary system, domestic price levels will vary in proportion to local note issues. The changes in the relative prices of the various currencies will equilibrate the demands for each currency with its supplies, and there is no need for gold flows to equilibrate the intercountry distribution of gold. The concept of the balance of payments loses much of its force, and international mobility of financial capital is necessary to let the value of exports (denominated in the home country’s currency) depart from the value of imports. The exchange rate moves to clear the markets for money and exports (or equivalently, imports): foreign exchange is required to buy imports, and as locals try to buy more and more imports, they bid down the price of their own currency in terms of the currency in which the imports

are denominated. Once internationally tradable financial capital enters the scene, the picture gets considerably more complicated, and we will defer such complications. Nonetheless, the monetary authority can control the domestic money supply under floating rates, but at the cost of being unable to control its currency’s value in terms of foreign currencies. It can control its domestic price level through its choice of money supply. In a purely metallic monetary system, the flexible exchange rate model could represent changes in the metal content of different countries’ issues, although the international (world) price of the metal might (or might not) remain unchanged.

9.5.6 *Seigniorage: making money by issuing money*

With metallic currencies, whoever mints coins from bullion can make a profit by charging more for the money than it costs. This is just like making a profit on the production and sale of any other good but with some critical exceptions. First, in most lines of production, competition keeps economic profits pretty close to zero, allowing only the competitive (“normal”) rate of return on capital. When governments mint coins, they frequently establish a monopoly for themselves, which gives them some power to make an economic profit from the minting. Second, charging more than it costs can take two forms: charging more for minting than the factor costs of the minting; and putting less metal in the coin than the established price of the metal would seem to prescribe for a coin of any given value – that is, short-changing the coin weight. The government can keep the “extra” metal, converting it into coins that it uses for its own purchases, as long as users accept them at face value. This conversion of the mint’s surplus from coinage into income for the government is called seigniorage.

However, if the value of the metal in a coin, at the government-fixed price of the metal, exceeded the face value of the coin, people would melt down the coin for the bullion, and the coins would disappear from circulation. They would be replaced

by bullion (dumps) if no other coins appeared to take their place.

As we have described them, these sources of seigniorage are stationary: if the weight of the coin stays the same fraction of the face value of the unit of account over some lengthy period of time, the coins will be accepted at face value without hesitation by people using them for transactions or as stores of value. As long as the increase in circulating coinage parallels the growth in the economy's production, the price level denominated in the coin's unit of account will stay roughly stable. However, if the precious metal content of the coinage is continually reduced, *and* the monetary authority mints the extra metal it shaved off the debased coinage into additional coins, the price level will gradually rise in proportion to the expansion of the coinage beyond the growth rate of real output in the economy. This inflationary price increase transfers purchasing power with new issues of coinage to the government from the public accepting the money. Contemporary governments, using fiat money of course, raise from 1–2% of their budgets among the more stable-price-level countries to as much as 12–15% for short periods of time in countries with chronic, high rates of inflation. This is what is colloquially called “the inflation tax.” It is indeed a tax. If a country exports its money through balance of payments deficits, some of the inflation seigniorage will be paid by foreigners, the balance being paid by domestic residents holding the currency.

9.5.7 *Bimetallism*

Bimetallism is the use of two or more metals simultaneously in a single country's currency. Generally the government fixes the prices of the metals in relation to one another. As long as the supplies of the two metals grow in pretty much the same proportion, a bimetallic standard can be stable. Although the Roman bimetallic system lasted quite some time (Goldsmith 1987, 36–40), more recent bimetallic standards have succumbed to Gresham's Law (cheap money drives out dear) more quickly. If either the market prices of the metals depart from the official

government parity or one of the coins is debased differentially relative to the other, it will pay to withdraw the coin with the greater intrinsic metal value from circulation, melt it down, and sell the bullion in terms of the other coin.

If coins of different metals circulate side by side at variable, rather than fixed, exchange rates, they can coexist indefinitely (Velde and Weber 2000). The floating (or flexible) exchange rates will be determined by the relative supplies of the different metals (assuming that the coins' services are perfect substitutes on the demand side, which might not hold if one of the metals had a long-term inflationary trend in supply). Demand for money changers will arise from the need (demand) to continually assess the exchange rates. If multiple coins of different standards circulate coextensively, the demand for money changers will be reinforced. These are real costs, and an economy could produce more consumable goods and services if the transactions costs conducted by the money changers could be avoided by unifying the monetary standard. The same conclusion holds for the substitution of a cheap, fiat currency such as paper for a more expensive metallic currency in general.

9.6 Inflation

In section 9.3.2 we described what inflation was but stopped short of theories of what causes it, exactly how it happens, or what it causes itself. We turn to those subjects here, now that we have discussed both the demand for and the supply of money. Inflation is the rate of change of the price or value of money, and it is difficult to speak economically about the change in the value of an object without referring to the conditions of the object's demand and supply. While we refer to inflation in this section, the case of deflation, or a declining price level is largely symmetric.

Much of the contemporary analysis of inflation has been directed at fairly high rates of inflation – say, 10% per year and above – but the concerns exist at lower levels as well, although we are reluctant to draw any particular lines below which inflation would not cause problems of one

sort or another. The ancient economies of the Mediterranean and Aegean, with their commodity currencies and unknown augmentations of bank money beyond metal coins, are unlikely to have experienced many episodes of 20–30% inflation per year, with a variance around such an average of, say, plus or minus 10 percentage points. The inflation tax is unlikely to pick up much revenue for a government whose currency expansion is letting the price level rise at 1% or less per year. Some recent price-level-change episodes that may be reasonable comparisons to ancient Mediterranean events are the world-wide inflation following the California gold discovery and rush of 1848–1855 and that in Australia of 1841–1853, the preceding decade of declining prices, particularly in the United States, and the extended period of deflation in the United States from just after the end of the Civil War until nearly the end of the century. While none of these inflationary or deflationary episodes reached the pace of inflation reached in the late 1970s in many Atlantic Community countries (15–20%), they did raise overall price levels throughout the world by 25–30% over a decade, and the gold rushes had dramatic local effects near the mining regions, with prices rising by 100–150% within two to three years. The possible differences in ancient and modern circumstances notwithstanding, we present a few salient analytical points about where inflation “comes from,” how it operates, and what it does.

9.6.1 *Causes of inflation*

A rallying cry of monetarists in recent decades has been that inflation is always and everywhere a monetary phenomenon. This position is in contrast to two other theories of inflation that gained popularity in the 1950s and 1960s, and retain some influence, if not particularly much credibility, today: the cost-push and demand-pull models of inflation. The cost-push explanation of inflation proposes that factor costs, particularly increasing wage demands of labor, raise production costs. Producers consequently raise product prices, which reduces the real wages

that rose in the first place, and so on, resulting in a “wage-price spiral.” According to the demand-pull mechanism, aggregate demand for goods grows by more than can be accommodated by production capacity, and goods prices rather than outputs, increase to clear the markets for goods. The higher prices of consumption goods reduce real income, resulting in pressures for wage increases, and we once again have a general, inflationary rise of all prices.

Neither model tells where the extra money comes from. If the money supply remains the same (or in a context of growth, if the money supply is growing at the same rate as aggregate output), and the demand for money remains unchanged (that is, its velocity remains constant), increases in some prices, either of products or factors, require decreases in other products. The price increases to which the cost-push and demand-pull models appeal are relative, not absolute, price changes. There is wide acceptance of this point today.

We are left with a supply-demand framework for studying the change in the price of money. This growth rate, as we noted early in this chapter, is inflation. Either the demand for money changes, or its supply changes. Consider the demand first, beginning from a situation of zero inflation. Money demand, at its simplest, is a function of the real interest rates on alternative assets and permanent income; since the current inflation rate is zero, we have identified nothing that would lead the expected inflation rate to differ, so the value of that argument is zero. Similarly with expected changes in the real interest rates: we have excluded current or anticipated inflation as reasons for departures of those rates from their current levels. That leaves us with “real” expectations of interest rate changes: any expected changes in real interest rates must come from expectations of real capital productivity changes – technological changes, or some such. Why don’t we ignore those potential sources of demand change for the moment, since we’d effectively be pulling something out of a hat to suppose that they had nonzero values in order to kick off an inflationary change in money demand.

We're left with unanticipated changes in real interest rates or permanent income as a source of change in money demand, and there are no general, theoretical grounds for giving ourselves such exogenous changes. We could appeal to war or pestilence as empirical events that could affect productivity so as to affect real interest rates, aggregate productive capacity, or both. If these grim-reaper events raised real interest rates and lowered productive capacity (the increase in real interest rates would tend to depress permanent income, or wealth, independently), they would depress the demand for money. People would get rid of some of their money holdings by spending on either goods or other assets. Even if the quantity of goods in the economy remained constant, more money chasing the same quantity of goods would raise the general level of prices; we appealed to events that reduced the quantity of goods, which would reinforce the price effect of the increased spending. Of course, we just appealed to war and pestilence to reduce the productive capacity of the economy, and if we look back to the transactions or income equation, $MY = PT$ or $MV = PY$, we see that we can't appeal to a money demand shift from such a source without a simultaneous shift in Y or T – the supply of goods. So this reduction in money demand is accompanied by a reduction in things to spend it on, and the quantity equations unambiguously tell us that the price level must rise as long as the nominal quantity of money remains unchanged. If our monetary system contains bank (deposit) money as well as gold currency, the rise in interest rates would encourage bankers to extend more loans, drawing down their reserve ratios; if the interest rate rise extended to the deposit rate as well, whether the net effect on bank money was positive or negative would depend to a great extent on what happened to the spread. If the spread remained the same, the profitability of loans would not change, although a rearrangement of balance sheet composition might have some minor effect in either direction. The simplest presumption is that money supply remains unchanged. The conclusion from this discussion is that a reduction in the demand for money, with

the supply of money unchanged, would raise the price level.

It is easy to see that an exogenous increase in the supply of money (one not induced by a change in the interest affecting the supply of bank money), with the demand for money unchanged, would increase the price level. An exogenous reduction in the quantity of goods in an economy, with no reduction in the quantity of money or in the demand for money, would have the same effect. Such a supposition may seem farfetched, but a tribute exaction, in real terms – wheat, nonprecious metals wherewith to make weapons with which to exact more tribute, people as slaves, and so on; but let's assume not in money terms for the moment – would have exactly this effect. A reparations payment in money terms would be deflationary, as long as only a country's money stock was removed and not real goods as well.

9.6.2 *Mechanisms of inflation*

The immediately preceding discussion has focused more on the difference in equilibrium price levels at different ratios of money stock to real goods, with a period of changing prices implicitly necessary to get from one price level to another. The demand for money is a stock demand in its portfolio component, as well as a flow demand related to the transactions uses of money. There is no reason to expect the adjustment of any stock to changed conditions to be instantaneous. The disposition of unwanted money balances could be expected to take time.

Suppose people all of a sudden find themselves holding a larger number of gold coins than they had previously held. Their real money balances have increased, because as long as they hold them, there's no reason for prices to increase. But, assuming they were already holding their equilibrium (that is, desired) level of real balances prior to getting the extra coins, they now have larger real balances than they demand. With an unchanged demand for real balances, they'll start getting rid of the extra coins. Spending the extra coins is a direct method, and one

that we discussed above in section 9.4.4. We also discussed a credit market mechanism in that section: the influx of money causes the interest rate on loans to drop below the rate of return on real capital, the flow of saving lagging behind the demand for investment a while, the demand for investment goods driving up their prices and the price level. The return of saving and investment to equality drives the loan and real capital returns back into equality and the rate of change of the price level (the inflation rate) back to zero at the new, higher price level.

Figures 9.4 and 9.5 show two hypothetical time patterns of the change in the quantity of money and the change in the price level. In contrast to Figure 9.2, which may look similar, there is some causality between the changes in the quantity of money and the change in the price level and the inflation rate in these two figures. Figure 9.4 shows a period of constant money stock and stable prices, that is, a zero inflation rate, from time t_0 to time t_1 . At time t_1 , the money stock increases instantaneously, but price changes lag behind a bit, not beginning to rise until time t_2 , when they begin to rise smoothly, with a positive but declining inflation rate, toward their new, equilibrium level at time t_3 . After t_3 , the inflation rate is again zero. In Figure 9.5, we offer a different growth path for the money stock. Beginning at time t_1 , the growth rate of the money stock increases from zero to some positive rate; by time

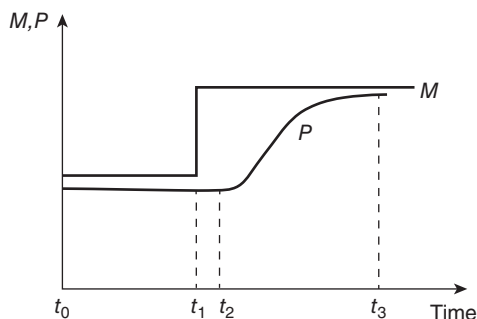


Figure 9.4 The time path of a money supply and the corresponding price level: instantaneous increase in the money supply.

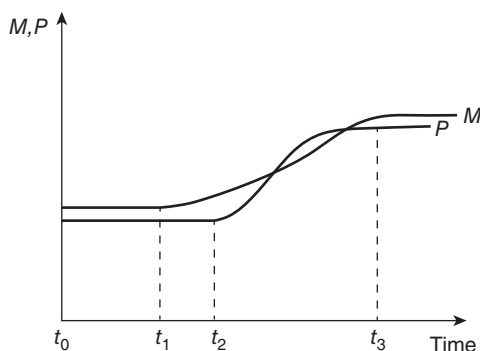


Figure 9.5 The time path of a money supply and the corresponding price level: gradual increase in the money supply.

t_3 , the growth rate of money has fallen back to zero, and the stock of money remains constant at its higher level. In this case as well, prices do not begin to respond to the growth in the money stock until time t_2 , but in this case, we have a short-term inflation rate greater than the growth rate in the money stock (we could appeal to a short-term depression in the demand for money, say because of an increase in the real interest rate). As drawn, the price level temporarily overshoots its long-term level by a small amount before it settles down to that level by time t_3 .

9.6.3 Consequences of inflation

When contemplating the consequences, or costs, of inflation, it is useful to distinguish between anticipated and unanticipated inflation. The former can be incorporated into people's plans at some, possibly moderate, cost. The latter disrupts plans, imposing potentially substantial costs, although this needs to be qualified. Changes in monetary policy can change the rate of inflation by altering the rate of growth of the money supply or by changing the relative cost of holding or using money, thus getting people to either spend a constant stock more quickly (inflationary) or slow down their spending (deflationary). The intention behind the use of monetary policy used as stabilization policy¹⁸ frequently is to get private

agents to do something they aren't doing under present circumstances. Otherwise it can be used to get something for the government on the hope that private agents will keep doing what they've been doing long enough for the monetary authority to skim off some income or wealth from those agents. But to the extent that people in an economy are able to anticipate short-term changes in monetary policy, they can be expected to adjust their own planned actions to incorporate it, thus partially or fully neutralizing the effect of the anticipated policy change. This discussion takes us slightly ahead of our storyline since monetary policy is the subject of the next section, but it fits naturally with the distinction between anticipated and unanticipated inflation and the consequences of monetary surprises on the "real" side of the economy. Unanticipated inflation in these ancient societies is at least as likely to have come from disruptions to the real side of the economy (wars or rebellions at home or abroad, weather, disease and pestilence) as to, say, new gold or silver strikes, which would have acted more slowly and probably with the awareness of the public. The existence of unanticipated changes in inflation rates for these times and places should not be ruled out.

It is also useful to distinguish between private and social costs of inflation. The inflation tax is a redistribution of income from holders of money to the government, and in that sense is not a net cost to the economy. To the extent that foreigners hold a country's currency, there is indeed a loss to the foreign country from inflation in the foreign currency they hold (but gains from deflation!) and a gain to the country issuing the inflating currency. The social loss to the country issuing the currency is similar to the deadweight loss of taxation discussed in Chapter 6. Figure 9.6 draws the demand for money as a proportion of income, M/Y (the ratio of real balances to real income), as a function of the inflation rate, $\dot{P}/P$. With a zero rate of inflation, the proportion of money to income Od would be held. With an inflation rate equal to a on the vertical axis, that proportion would fall to Oc . The area of the rectangle $Oabc$ is the proportion of real income that money holders

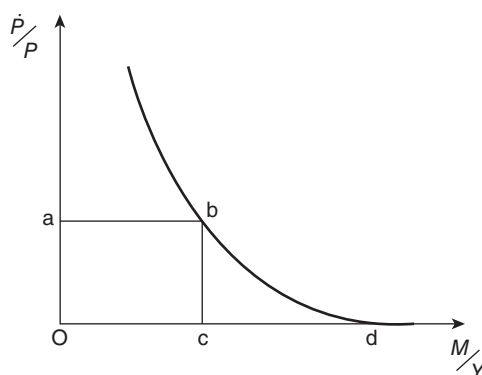


Figure 9.6 Demand for money as a function of the inflation rate.

are obliged to accumulate in the form of money just to keep their real balances intact at the desired (demanded) level, which is the inflation tax as a proportion of real income. The triangle bcd is the loss that society incurs in its efforts to substitute away from money because of the higher holding cost imposed by the inflation rate. These substitutions can take the form of shortened payments contracts, holding stocks of goods rather than stocks of money, and so on. The rate of return on the resources used in these activities will be bid up to the rate of inflation.

Not all the consequences of inflation, at least mild inflation, need be unambiguously undesirable. In a growing economy, a mild rate of inflation, by reducing the rate of return on cash, can induce a substitution from assets held as cash balances to real capital. This increases capital accumulation and leads to growth in the real wage and real income. There is some empirical evidence that mild rates of inflation are associated with higher economic growth rates. The tendency of some countries toward chronic, even galloping inflation seems to involve a causation from slower growth to the practice of inflationary finance by governments using fiat, rather than commodity, money.

Both inflation and deflation redistribute income. Inflation reduces the obligations of debtors while deflation increases them. Indexation can

limit such redistribution, but requires potentially extensive measurement. When a government owes considerable fixed-price claims to its citizens, it can reduce the value of that debt by inflating the currency. Despite a few efforts to identify inflationary income redistribution as operating in the net direction of one social class or another (other than between debtors and creditors in general), there do not seem to have been any such easily identifiable trends.

9.7 Monetary Policy

Some writers have been inclined to see little scope for monetary policy in the economies of the ancient Mediterranean region. Certainly there was less scope for deliberate use of the monetary system to achieve indirect economic goals than there is in most countries today, with their considerably more intricate financial and monetary institutions. The intentionality of some of the ancient monetary manipulations may be questioned on the grounds of lesser ancient understanding than contemporary (which still has ample room for improvement). Nonetheless, the actions that governments could have taken *vis-à-vis* their monetary systems may be worth considering just to see what might turn up under closer scrutiny. Nonetheless, many of the policy choices we describe below certainly did exist. We do not claim that the ancient understanding of the full array of consequences of some of these choices was particularly well fleshed out,¹⁹ but some of the consequences, even some subtle ones, of some choices were well understood. Of course, as the recent history of banking demonstrates amply, even misunderstandings of consequences offer interesting grist for the analytical mills. In other instances, in which choices may have been made with little thought given to consequences, consequences would have occurred just the same: they don't wait for understanding. For instance, usury laws have been enacted for religious reasons, with apparently little thought to their real effects. Consequently, we give accounts of many of the choices that may have been made and, if they were, would have

affected money and related activity in one fashion or another.

It is widely thought, although without consensus, unlikely that ancient governments regularly attempted to use short-term changes in monetary policy for stabilization purposes – although interest rate limits and partial debt cancellations had been implemented by the Roman Republican government in the fourth-century B.C.E., and the *lex Cornelia Pompeia* of 88 B.C.E. appears to have been an effort by the government to alleviate a financial panic (Barlow 1980, 212–215); and Lo Cascio (1981) makes the cases for the Roman Republican government loosening the money supply on a number of occasions of financial difficulty during the Late Republic and for continual Roman governmental intervention to maintain the official ratio between the gold aureus and the silver denarius during the Imperial period. Limitation of the purposiveness of ancient monetary policies to only providing a public service in facilitating transactions and getting a return in seigniorage may underestimate those governments' goal-motivated, monetary activism.

9.7.1 *The players and their motives*

The question here is, "Who gets to make monetary policy?" The answer to this question depends in part on what we include in monetary policy. Some policy decisions involve long-term, structural choices such as the choice of standard; others address choices that may be changed more frequently with low cost, such as the rate of increase of the money supply. Yet others fall in between on this spectrum of frequency/cost of change, such as banking regulation and law.

In contemporary industrialized countries, authority and responsibility for choices affecting the monetary system are divided among political bodies (executives and parliaments or congresses), bureaucratic political agencies with varying degrees of responsiveness to current political demands (finance ministries or treasuries), and relatively independent central banks. The strictly political bodies face a host of demands on their authorities, and they are

in positions to respond to questions of income distribution among domestic constituents. The Silver Controversy in the United States in the second half of the nineteenth century and the first decade or so of the twentieth, dealt with whether the U.S. monetary system would be fully bimetallic or remain effectively on a gold standard. The monetary consequence at issue was whether a monetary system would be instituted that would let prices rise by more (gold plus silver) or less (gold only), which would affect interest rates on loans. The distributional issue was perceived to be whether greater income growth of agriculturalists and small businesses would be fostered at possibly slower growth of income for larger industrialists. The matter was addressed in the principal national political institution, the Congress. In some of the states of the first millennium B.C.E. and first millennium C.E., such choices probably would have remained political, and the locations of their decisions would have been the kings, princes, and emperors and their advisers, many of whom would have had personal interests in the decisions (which doesn't necessarily distinguish them from contemporary legislators).

The principal players in contemporary monetary policy are central banks and finance ministries or treasury departments, with the longer term ground rules set by legislatures. Many of the longer term policies affecting the monetary system are rules for banks to follow; these typically are overseen and enforced by regulatory agencies, either independent of the central bank and treasury or as subagencies of those institutions. It is not difficult to find the equivalent of treasuries or finance ministries in many of the ancient economies of the Mediterranean area, but for central banks and bank regulatory agencies we must look for particular duties scattered across other institutions.

Independent central banks, which are critical players in contemporary monetary systems, are relatively recent institutional developments. Some of the choices they make today were not available, technologically, to be made in, say,

1000 B.C.E., or 150 C.E., but others either certainly or probably were. The decisions that *were* available to be made, clearly were made by some other institutions. Together with the placement of ancient monetary responsibilities in different institutions from those extant today would have come different understandings of how monetary relationships worked as well as different mechanisms for determination of goals.

9.7.2 *Choice of monetary standard*

A particularly long-term decision is the choice of metal or metals to be used as the monetary standard. The absence of any particular choice is a decision in favor of accepting whatever the markets will accept. Some countries' monetary standards could actually be other countries' currencies, which yields the seigniorage to the issuing country but retains the transactions benefits for all users. Small countries would have been more likely to use foreign currencies than would large countries; the fixed costs of a mint and the trouble of importing steady supplies of precious metals may have been not worth the trouble.

9.7.3 *Influencing the supply of money*

Perhaps the primary decision regarding money supply that an ancient government could have made was whether to enter the money supply industry or leave it to private parties. If a government left the money supply to private agents, it still could have imposed some regulation on those agents' activities. For instance, taxation of money changers could have affected the quantity of money by making its supply more costly, if not by much. Whether taxation of private monetary agents would have been a deliberate effort to affect their production decisions is another matter.

Assuming the government exercised authority over the monetary system, decisions regarding free or controlled coinage at mints would have given governments more or less control over the growth of their money supplies. If free coinage,

as we described it above, were adopted, the government would largely abdicate control over the growth of the money supply. Controlled coinage would have given control over supply to the government, at least to the extent that volumes mined were not constraining.

Currency debasement – reducing the metal content of coinage – is a policy option equivalent to an inflationary expansion of the money supply, but only if the government takes advantage of the extra metal saved from the debased coins to mint other debased coins. If the same quantity of coins were kept in circulation, the only difference being that the new ones minted as old ones were retired had a lower metal content, there would be no domestic pressure for inflation. In fact, there would be a saving in the cost of maintaining the money supply since fewer resources would be tied up in the circulating medium. The reduction in the metal content would be equivalent to going to a lower commodity “backing,” the extreme limit of which is contemporary fiat currency, which has no commodity backing at all at the moment. International connections with other metallic currencies could change this result: the metal content of the currencies, at constant relative prices of the precious metals, could create a flexible exchange rate, which would raise prices in the debasing country.

To the extent that governments were in the mining business, particularly for precious metals, the allocation of resources to exploitation of known deposits and to exploration for new deposits would have influenced the money supply. Again, this would have been a relatively indirect monetary policy, and certainly not a familiar, contemporary policy “instrument.”

Placing restraints on interest rates, either asked by banks or by private individuals, would have affected the equivalent of the M1 money supply. Usury restrictions such as interest-rate ceilings on various types of deposits limit the funds available to lend. At the same time, however, by placing an upper limit on the lower end of the loan-deposit rate spread, they can increase the profitability of loans, given a particular level of deposits, which could encourage banks to operate with a lower reserve ratio (with a minimum reserve

requirement, a ceiling, or an increase in it, could reduce excess reserves). The effects of a change in the interest rate ceiling on some category of deposits (say, demand deposits) would be more intricate with a wider array of assets and liabilities in the balance sheets of banks than was probably the case for ancient banks, but at any rate, the net effect, whether expansionary or contractionary, on the money supply probably would be small. If private banks were to turn to the equivalent of the finance ministry or treasury for a short-term loan, the interest rate required by the government agency would be the equivalent of the contemporary, central bank discount rate, increases in which are intended to tighten the money supply, and decreases to loosen. Using interest rates to adjust the money supply makes the money supply an endogenous variable – responding to interest rate changes – rather than an actual “control variable” of the monetary authority or government such as would be conferred by direct control over the monetary base.

A minimum reserve requirement against deposits imposed on banks restricts the expansiveness of any particular volume of deposits. Raising the reserve requirement is a potent means of tightening the money supply (decreasing its rate of growth, possibly even its quantity), and conversely for a reduction in reserve requirements.

The simple distinction between monetary and fiscal policies – the latter being spending policies of government – blurs the frequent interaction between the implementation of a fiscal policy and the incidental exercise of a monetary policy. Government spending puts additional money into circulation. How the government obtains the funds and how it injects them into the economy both can affect the growth of the money supply.

It has been rather traditionally assumed that the scope for monetary policy in most of these ancient economies was quite restricted. In comparison to contemporary practice, the ancient opportunities, both those extant and those actively taken, must seem marginal, but closer examination of the ancient textual evidence might yield substantial rewards, as examination of the Roman record seems to have opened up.

9.7.4 *Influencing the demand for money*

Most deliberate monetary policy operates on the supply of money. While manipulation of interest rates will affect the supply of bank money – assuming banks are creating money, which must remain a point for case-by-case demonstration – interest rates also affect the demand for money. Governments, just like the producers of any good, have an interest in consumers' demand for their money. Instability and unpredictability of value of a currency will reduce the demand for it, so we may consider deliberate actions to maintain the quality of a currency as demand-side policies. These would include the quality control in minting (how close to a uniform weight are different individual coins?) and the reputation for frequency of debasement.

9.7.5 *International monetary policies*

In a fixed-rate system of international exchanges, the exchange rate a government establishes *vis-à-vis* foreign currencies is a key international policy that is subject to change, if, ideally, infrequently. Exchange rates cannot be chosen and successfully defended (by buying and selling the domestic currency to maintain its price) if they are chosen arbitrarily. Neither can changing the exchange rate be done arbitrarily: only when the fundamental demand for and supply of the currency has changed – and then the new rate must be picked carefully and accurately to reflect the new equilibrium price or it will not be defensible.

Whether a country will permit foreign currencies to circulate within its borders is a policy choice that governments have. Such an allowance

is more likely when metallic currencies dominate money supplies rather than paper notes. As a government can earn seigniorage on its own currency issues but not on foreign issues, there is some incentive to restrict foreign circulating media.

9.8 **Suggestions for Using the Material of this Chapter**

The concepts surrounding money and banking will inform ancient historians and philologists primarily. Interpretation of coin hoards may involve archaeologists as much as the historians and philologists.

Of primary importance, comparing prices at different dates requires adjustment for price-level differences. Comparisons without price-level adjustments can be meaningless. Granted, a number of ancient economies experienced lengthy periods of great price stability – inflation rates in the neighborhood of 0.1% a year. However, a burst of inflation of 1% a year for 50 years would raise prices by about 60%. Comparison of nominal prices (that is, unadjusted for even gradual price-level change) at dates a half-century apart will produce meaningless results.

Similarly with interest rates. Nominal interest rates during periods of inflation require adjustment with the inflation rate to determine the real rate. Recall that the real interest rate is the nominal (or stated) rate minus the rate of (expected) inflation. To compare interest rates at different dates and at different places (using different currencies), the expected inflation rate in the two places, at the different dates, must be subtracted from stated inflation rates to make meaningful comparisons.

References

-
- Andreau, Jean. 1999. *Banking and Business in the Roman World*. Translated by Janet Lloyd. Cambridge: Cambridge University Press.
- Barlow, Charles T. 1980. "The Roman Government and the Roman Economy, 92 – 80 B.C." *American Journal of Philology* 101: 202–219.
- Baumol, William J. 1952. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Quarterly Journal of Economics* 66: 545–556.
- Brewer, Douglas J., and Emily Teeter. 1999. *Egypt and the Egyptians*. Cambridge: Cambridge University Press.

- Cagan, Phillip. 1956. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago IL: University of Chicago Press, pp. 25–117.
- Cohen, Edward E. 1992. *Athenian Economy and Society; A Banking Perspective*. Princeton NJ: Princeton University Press.
- Cohen, Edward E. 2008. "The Elasticity of the Money-Supply at Athens." In *The Monetary Systems of the Greeks and Romans*, edited by W.V. Harris. Oxford: Oxford University Press, pp. 66–83.
- De Graef, Katrien. 2008. "Giving a Loan is like Making Love . . ." In *Pistoia dia tèn Technèn; Bankers, Loans and Archives in the Ancient World. Studies in Honour of Raymond Bogaert*, edited by Koenraad Verboven, Katrijn Vandorpe and Véronique Chankowski. Leuven: Peeters, pp. 3–15.
- Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91: 401–419.
- Dowd, Kevin. 1990. "The Value of Time and the Transactions Demand for Money." *Journal of Money, Credit and Banking* 22: 51–64.
- Dutton, Dean S., and William P. Gramm. 1973. "Transactions Costs, the Wage Rate and the Demand for Money." *American Economic Review* 63: 652–665.
- Friedman, Milton. 1956. "The Quantity Theory of Money: A Restatement." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago IL: University of Chicago Press, pp. 3–24.
- Friedman, Milton, and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867–1960*. Princeton NJ: Princeton University Press.
- Goldsmith, Raymond W. 1987. *Premodern Financial Systems; A Historical Comparative Study*. Cambridge: Cambridge University Press.
- Gorton, Gary, and Andrew Winton. 2003. "Financial Intermediation." In *Handbook of the Economics of Finance: Corporate Finance, Volume 1A*, edited by George M. Constantinides, Milton Harris, and René M. Stulz. Amsterdam: Elsevier, pp. 431–552.
- Gurley, John G., and Edward S. Shaw. 1960. *Money in a Theory of Finance*. Washington, D.C.: The Brookings Institution.
- Hamilton, Earl J. 1934. *American Treasure and the Price Revolution in Spain, 1501–1650*. Cambridge MA: Harvard University Press.
- Harris, W.V. 2008. "The Nature of Roman Money." In *The Monetary Systems of the Greeks and Romans*, edited by W.V. Harris. Oxford: Oxford University Press, pp. 174–207.
- Karni, Edi. 1973. "The Transactions Demand for Cash: Incorporation of the Value of Time into the Inventory Approach." *Journal of Political Economy* 81: 1216–1225.
- Kemp, Barry J. 2006. *Ancient Egypt; Anatomy of a Civilization*, 2nd edn. London: Routledge.
- Kim, Henry S. 2001. "Archaic Coinage as Evidence for the Use of Money." In *Money and Its Uses in the Ancient Greek World: Approaches to the Economics of Ancient Greece*, edited by Andrew Meadows and Kirsty Shipton. Oxford: Oxford University Press, pp. 7–21.
- Kim, Henry S. 2002. "Small Change and the Moneyed Economy." In *Money, Labour and Land in Ancient Greece*, edited by Paul Cartledge, Edward E. Cohen, and Lin Foxhall. London: Routledge, pp. 44–51.
- Lerner, Eugene M. 1956. "Inflation in the Confederacy." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago IL: University of Chicago Press, pp. 163–175.
- Lesko, Barbara S. 1994. "Rank, Roles, and Rights." In *Pharaoh's Workers: The Villagers of Deir el Medina*, edited by Leonard H. Lesko. Ithaca NY: Cornell University Press, pp. 15–39.
- Lo Cascio, E. 1981. "State and Coinage in the Late Republic and Early Empire." *Journal of Roman Studies* 71: 76–86.
- Lucas, Robert E., Jr. 1980. "Rules, Discretion and the Role of the Economic Advisor." In *Rational Expectations and Economic Policy*, edited by Stanley Fischer. Chicago IL: University of Chicago Press, pp. 199–210.
- McCallum, Bennett T., and Marvin S. Goodfriend. 1987. "Demand for Money: Theoretical Studies." In *The New Palgrave; A Dictionary of Economics*, Vol. 1, edited by John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan, 776–777.
- Millett, Paul. 1991. *Lending and Borrowing in Ancient Athens*. Cambridge: Cambridge University Press.
- Mitchell, Wesley Clair. 1903. *A History of the Greenbacks; With Special Reference to the Economic Consequences of their Issue: 1862–65*. Chicago IL: University of Chicago Press.
- Niehans, Jürg. 1978. *The Theory of Money*. Baltimore, MD: Johns Hopkins University Press.
- Rathbone, Dominic, and Peter Temin. 2008. "Financial Intermediation in First-Century AD Rome and Eighteenth Century England." In *Pistoia dia tèn Technèn; Bankers, Loans and Archives in the Ancient World. Studies in Honour of Raymond Bogaert*, edited by Koenraad Verboven, Katrijn Vandorpe and Véronique Chankowski. Leuven: Peeters, 371–410.

- Rockoff, Hugh. 1984. "Some Evidence on the Real Price of Gold, Its Costs of Production, and Commodity Prices." In *A Retrospective on the Classical Gold Standard, 1821–1931*, edited by Michael D. Bordo and Anna J. Schwartz. Chicago IL: University of Chicago Press, pp. 613–650.
- Schumpeter, Joseph A. 1954. *History of Economic Analysis*. New York: Oxford University Press.
- Snell, Daniel C. 1995. "Methods of Exchange and Coinage in Ancient Western Asia." In *Civilizations of the Ancient Near East*, Vol. III, edited by Jack Sasson. New York: Scribner's, pp. 1487–1497.
- Spiegel, Henry William. 1983. *The Growth of Economic Thought*, 2nd edn. Durham NC: Duke University Press.
- Temin, Peter. 2004. "Financial Intermediation in the Early Roman Empire." *Journal of Economic History* 64: 705–733.
- Temin, Peter. 2006. "The Economy of the Early Roman Empire." *Journal of Economic Perspectives* 20, 1: 133–151.
- Tobin, James. 1958. "Liquidity Preference as Behavior toward Risk." *Review of Economic Studies* 25: 65–86.
- Tobin, James, with Stephen S. Golub. 1998. *Money, Credit, and Capital*. Boston MA: Irwin McGraw-Hill.
- Velde, François R., and Warren E. Weber. 2000. "A Model of Bimetallism." *Journal of Political Economy* 108: 1210–1234.
- Verboven, Koenraad. 2009. "Currency, Bullion and Accounts. Monetary Modes in the Roman World." *Belgisch Tijdschrift voor Numismatiek en Zegelkunde / Revue Belge de Numismatique et de Sigillographie* 155: 91–121.
- Verboven, Koenraad, Katelijn Vandorpe and Véronique Chankowski. 2008. *Pistoia dia tèn Technèn; Bankers, Loans and Archives in the Ancient World. Studies in Honour of Raymond Bogaert*, edited by Koenraad Verboven. Leuven: Peeters.
- von Reden, Sitta. 2010. *Money in Classical Antiquity*. Cambridge: Cambridge University Press.

Suggested Readings

- Burger, Albert E. 1971. *The Money Supply Process*. Belmont CA: Wadsworth.
- Laidler, David E.W. 1977. *The Demand for Money: Theories and Evidence*, 2nd edn. New York: Harper & Row.
- Niehans, Jürg. 1978. *The Theory of Money*. Baltimore, MD: Johns Hopkins University Press.

Notes

- 1 For deposit banking, and consequently "bank money" in Classical and Hellenistic Athens, see Cohen (1992, 11–18, 114–121).
- 2 I'm trying to walk a fine line here with the language. It is one thing to compare unchanged and unchanging situations – for example, one level of M with another level of M , with all the other variables keeping the same values – and quite another to say that the very same system that had one configuration of values for these four variables now has one of them change in magnitude. The latter is more intricate, because we may not be able simply to assume that two of the other three variables keep their original magnitudes while the third alone changes in response to the initial change we posit in the fourth. This probably sounds rather abstract. Suppose we said that, at time zero, the quantity of money, M , increased by 10%, and we wanted to know what would happen to the price level. We could very well just say that T stays the same – possibly because real productive technologies are unaltered and unaffected by the change in money; this might or might not be a particularly interesting assumption to make, but it violates no behavioral beliefs we hold about how "the system works." However, it would be more difficult to justify the assumption that V remains the same; we haven't considered the behavioral determinants of V yet, but to get slightly ahead of our story, there is good reason to suspect that, in the course of real time, V might respond somewhat to the change in P induced by the hypothesized change in M . Economists use a loose adaptation of the *ceteris paribus* phrase, for which they are notorious, to describe such a circumstance as one in which the *ceteris* aren't *paribus*. Welcome to some of the intricacies of monetary theory!
- 3 Ancient monetary systems offer challenges to constructing time series of even individual prices,

much less measures of the aggregate level of prices. Von Reden (2010, 148–151) offers an instructive example of converting different currencies into compatible units.

- 4 See Spiegel (1983, 40) for an admittedly vague report on Egyptian wheat prices in the first to third centuries C.E.
- 5 It is worth citing another view of Schumpeter's at this point (1954, 64 n. 8): "Observe: I am not arguing against Aristotle's ideal of life or against any particular value judgments of his. Still less am I arguing for glorification of economic activity. On the contrary, I applaud the philosopher for having refused to identify rational behavior with the hunt for wealth. All I want to establish is that Aristotle, who in political matters was so alive to the necessity of analyzing and fact-finding as a preliminary to judging, never seems to have bothered about this preliminary in 'purely' economic matters except in the matters touching value, price, and money. For instance, the fundamental difference he finds between the trader's and the producer's gains is essentially preanalytic. This fact has nothing to do with the other fact that he disapproved of the former and approved of the latter."
- 6 Earlier versions of these theories commonly assumed that money earned no interest. This is commonly a well justified assumption but is by no means necessarily true, as the recent experience of interest-earning demand (checking) deposits demonstrates. This causes no problem in defining the opportunity cost of holding money, since the interest cost can simply be restated as the difference between the interest rate that money (checking deposits) can earn and the rate available on other assets.
- 7 More precisely, the discounted marginal value of a stock of money holdings – the value of an increment to the stock – which is equivalent to the percentage difference in the flow values of money at two time periods, is equal to the interest rate offered on money.
- 8 Keynes became noted for his excessively cute statement that, "In the long run, we're all dead." This was part of Keynes' rhetoric for promoting his new theory of aggregate income and employment determination. He knew very well himself, as we saw ourselves in Chapter 2, that we are always in some long run: every short-run equilibrium (by which we mean a configuration of values of a particular set of variables) is a position on a particular long-run curve. Do not be distracted by the rhetorical flourish. It was intended to be distracting.
- 9 A well developed example of this approach is Niehans (1978, Chapter 4); also see McCallum and Goodfriend, (1987, 776–777) for a summary account.
- 10 This section is adapted from Niehans (1978, Chapter 8).
- 11 For an analysis of the role of asymmetric information and real production in banks' issuance of demand deposits, and the consequent sensitivity of banks to the unstable equilibrium known as the "run," see Diamond and Dybvig (1983).
- 12 This section is based on Niehans (1978, Chapter 9). The treatment of the banking firm in Tobin and Golub (1998, Chapter 7) offers considerable detail on bank portfolios and the costs and profitability of deposits, but it is also closely tailored to contemporary American banking regulation.
- 13 This analysis, and some of the others we have conducted in this chapter, have departed from the relatively standardized, "off-the-shelf" character of the models we have used for most of the price theory in the previous chapters. Those earlier models were more generalized and had more the character of building blocks – whatever you want to analyze, you use one form of them or another. We were able to avoid specifying much of the environment in which the actions took place. The models presented in this chapter, and to a great extent the ones I will present in subsequent chapters have more the character of specialized models; they are not the only possible representation of the interactions under study. In some cases, some economists would disagree with the particular formulation while others would endorse them. In other cases, most or all economists would agree with the general structure of the formulation although they might be inclined to implement it in different ways. These models, and the fact that we have little choice but to follow this course of analysis, highlight an important fact about economic analysis: generally the answers to specific questions aren't known in advance, and once some answers are produced, they are fairly (highly) circumstance-specific. There aren't a lot of answers that are both very general and very interesting. Whether economists consider this a virtue or a shortcoming of their science is less important than the fact that they consider this

characteristic to reflect the state of the world and human behavior. Doctrinaire answers are harder to deliver.

- 14 This section relies on the classic work in the field, Gurley and Shaw (1960, Chapters 4 and 6).
- 15 Gorton and Winton (2003, 431–552) focuses on the bank as a financial intermediary.
- 16 Temin (2004, 708) suggests that most investment in the early Imperial period was self-financing; he appears to have changed his mind since: cf. Rathbone and Temin (2008). Raymond W. Goldsmith (1987) by and large offers a view congruent with Temin (2004) of the Augustan Roman financial structure (43–50) and a similar view of the inconsequentiality of financial intermediation in Classical (Periclean) Athens (27–33), but in both instances relies unreliably on Finley. More recently, Cohen (2008, 76–81) has provided evidence of extensive commercial deposit banking (Temin’s financial intermediation) in Classical Athens.
- 17 To be discussed below; basically, the revenue earned by issuing money with face value greater than its cost or intrinsic value.
- 18 Stabilization of employment and income over a business cycle.
- 19 I hesitate to say “correct” because contemporary theoretical beliefs continue to rise and fall themselves; as far as the aptness of our understanding goes, Robert Lucas, who won the Nobel Prize in economics in 1996 for his contributions to macroeconomic theory, said it well in Lucas (1980, 209): “We’re way over our heads in this business [of giving policy advice].”

Labor

10.1 Applying Contemporary Labor Models to Ancient Behavior and Institutions

The student of ancient societies may be reticent when it comes to studying the work behavior of ancient people with theories and models that are used comfortably in the analysis of contemporary, industrialized labor markets. After all, to work in the typical Western country one needs a handful of papers proving age, citizenship, current residence, promises to pay taxes, various dimensions of health, educational achievement, previous work history, legal record if any, and probably other bits of information. The person, or more commonly the institution, hiring her frequently needs a small army of accountants, legal advisors, safety and comfort specialists, insurance and pension specialists, bookkeepers, and probably a half-dozen other types of specialist to stay on the right side of the law. Private individuals commonly find it legally difficult to hire nannies and housekeepers without running the risk of fines or jail time for failure to comply with the maze of government regulations. We refer routinely to compensation instead of something simple like pay, because of the combination of wage or salary, complicated overtime payment schedules, several types of insurance, even more types of pension

options, vacation and other leave options that are formalized into work contracts. How can an analytical framework that has been found useful to operate in such an environment, not to speak of one that has had many of its features designed specifically to address some of those features, possibly have any relevance to the efforts of labor 2000 to 4000 years ago?

Let's take a specific example of the potential difficulty. The theory of labor supply predicts how much time an individual will furnish to employers outside the household at various wages; by its very structure it presupposes an institutional setting in which the proportions of work within one's own household and work for others is very nearly the inverse of what it probably was in most times and places in the ancient Mediterranean civilizations and societies. What can be its pertinence? To pursue the empirical correspondences first, most societies outside the contemporary industrial world have offered opportunities for people to work at least some of their time outside the household. The contemporary labor supply models will apply to those situations. More compelling is the theoretical rationale: the same forces that influence the allocation of a person's time between activities in the household – either leisure or household work – and work outside the household influence

the allocation of time among activities within the household. The primary difference between intra-household and extra-household time allocations is that the productive value of time (you can call it the wage rate if that's not too offputting) is less likely to vary according to one's own time allocation when supplying labor outside the household than within the household, and the analytical structures (the models) needed to accommodate that interdependence are more complicated than what a simple, fixed wage calls for. The household production model was introduced in Chapter 3; it studies how people combine time within the household – time not spent working for others – and goods to produce final consumption. The goods could be either bought in markets or produced oneself with another of these household production functions. At one end of the spectrum of outside work opportunities lies the Robinson Crusoe case: the completely closed (autarkic) household that neither gets goods or services from outside nor supplies any of its own products or services outside its confines. The popularity of the Crusoe parable within the traditions of economic theorizing should be a warning to any who might be concerned that economic theory applies only to activities in markets. Despite all the issues that the household production model alone can illuminate, on the subject of labor supply it offers no insights beyond those yielded by the simpler model in which the supply of labor is represented by an individual's time crossing the invisible border between the household and the world beyond (we could call it the labor market, although there is no need to associate all the institutional complications of our opening paragraph with a market in labor). Many of the topics in "market" labor supply carry over, with greater complication, to the intrahousehold allocation of effort: responses of effort to labor productivity, to greater income and wealth, the timing of effort over the lifecycle, the influence of one household member's productivity on another's time allocation, and so on.

Labor is organized within and between two principal institutions in any society. Families or households supply labor, and firms or production enterprises by any other name demand labor.¹ Sometimes families are the demand institutions

as well, in which case they combine the institutions of both labor supply and labor demand. The family farm has been a widespread production unit until recent times in industrialized countries and retains major quantitative importance in many contemporary developing countries. The family production unit, probably most of them agricultural producers, surely was common in antiquity. Of course, large institutions such as temples and royal houses operated at various times in Egypt and Mesopotamia, probably in parallel to private households, possibly structurally akin to plantations in recent times.²

There may be an implicit, "default" view of ancient work behavior that everybody old enough worked from sunup to sundown all his or her life just to stay alive. If, or where, held, this would not be a particularly discriminating view. It is true that empirical evidence on the finer details of working life, and frequently even its major outlines, are obscure or lost, but the scholar's awareness of more discriminating concepts will aid in his or her thinking about ancient behavior – and may even assist in deriving more information from what evidence does remain. The theories and models of this chapter will deal with what might be considered some of the finer details about working life, as well as some major structural facts: the influences on hours of work, on the allocation of time across activities, on the time spent working for oneself and for others, on the size of remuneration (how much people were paid), on the units of activity for which labor was paid (for example, time versus piece rate), on the supply of effort versus the simple supply of time, on contractual relations between demanders and suppliers of labor, on working conditions – in other words, much of the richness of ancient, everyday life.

The subject of labor is inherently about people, as contrasted with, say, things, and the economics of labor targets the distinguishing characteristics of people and the differences in behavior among individuals. I have chosen to open the chapter (section 10.2) with a further discussion of human capital (introduced in Chapter 8), as the analytical foundation for studying differences among individuals and the effects of those differences on their behavior. So far, our treatment of

economics has focused on differences in external circumstances as the prime influence behind differences in behavior; this chapter enriches rather than replaces that view. Given a general level of knowledge in a society at a particular time, how was that knowledge distributed across individuals? How did the people who had it acquire it? How did they decide what knowledge to acquire themselves and what to leave to others? Knowledge is a powerful source of productivity. What are the institutions for acquiring human capital in a particular society: schools, employers, tutors, family, unions or guilds, the state? What influences the mix of institutions, what the institutions provide, their existence at some times and places and not others?

I have presented economics in a framework of supply and demand, and it is natural to treat labor in this fashion also. I follow the treatment of human capital with the theory of labor supply (section 10.3) based on the utility-maximization model, which underlay consumption theory in Chapter 3. This analysis centers on the family or household and leads into more extensive considerations of the family: household production in subsection 10.3.4, and intrahousehold allocation of resources, reproductive decisions, and production in the family enterprise in separate sections. Section 10.4 presents the theory of labor demand as the demand for a factor of production, derived from the demands for the products made with labor and other inputs.

People, the tasks to be done by them, and the conditions under which they are to be done are diverse. Agreements about methods of payment and levels of effort and diligence need to accommodate this diversity. What may look like tremendous standardization of labor contracts in antiquity – say, *X* quantity of grain per day for “everybody” for a period of 200 years of archaeological and textual evidence – may simply indicate that much of the agreement was unwritten and implicit. But the quantity of payment is only one component of a labor contract; length of time spent working, effort, periodicity, the costs of “getting to work,” the types of monitoring, penalties for breaking parts of the contract, and certainly others – all may vary.

Section 10.5 discusses issues in the contractual agreements that accommodate the diversity of people, technologies, and circumstances.

While much of the allocation of labor occurs while the family or individual doing the allocation remains in one location, migration allocates labor to different locations. While many of the great migrations of the past three centuries have had large voluntary components, refugee relocations and invasions have been prominent in either the evidence or suspicions in the ancient Mediterranean Basin. Section 10.6 addresses economic motivations for migration and economic consequences of migrations in which free choice can play the principal role or a relatively minor one.

The family, the subject of section 10.7, is a signally important institution in the economy. It is the source of labor supply and the location of consumption. Allocation of resources across individuals – the source of much equity or inequity within a society – occurs within the family. Child care is one important use of time – work and leisure being alternatives – and this fact links fertility to all other economic decisions, both consumption and production.

As I have noted several times even in this section, the quantitative importance of working for others by choice (that is, aside from slavery) until fairly recently was much smaller than it is today. The family was the principal production unit – effectively a firm. Labor use “on-farm” versus “off-farm” is an important analytical distinction, but either way, it’s still labor supply. The existence of landless labor implies that somebody is working for somebody else (except, of course, in the case of families of pure craftsmen, who would have possessed other items of capital stock than farm land). Section 10.8 treats labor use in the family enterprise, focusing on the influence on labor of imperfections in, or even the absence, of other markets, particularly intertemporal markets.

The chapter closes with a brief treatment of the economics of slavery. I dealt with slavery as a problem in capital theory in Chapter 8, and I retain that emphasis here but some of the labor-allocational concepts developed earlier in this chapter apply to topics in slavery.

Before embarking, it is important to note the multidimensionality of transactions in the labor market – or in transactions involving labor but occurring outside of markets. Many terms must be agreed upon: wages to be paid, the level of effort to be put into tasks, the range of tasks that the worker can be asked to do, the conditions of work, the duration of the contract, and with many of these terms varying according to the age, sex, training, and other characteristics, acquired and immutable, of workers. The basic analysis of labor supply and demand focuses on wages and hours of work supplied, abstracting from these complications of individual people and tasks. Consequently the explanatory and predictive power of the theories of labor supply and labor demand, in subsections 10.3 and 10.4, is limited – but still worth the effort. Most of the rest of this chapter is devoted to the analysis of these individualistic complications in the allocation of time and labor. One last note regarding this chapter's presentation: the analysis of many topics in labor requires keeping track of many details – different people, different time periods, different activities, arrays of goods and possible actions, and so forth. Consequently I frequently use the simplifying device of formulaic notations, but I take pains to explain what these notations mean and what the formulas are showing. The level of understanding required is consistent with the first five chapters of the text.

10.2 Human Capital

We can define human capital as all acquired characteristics of people that make them more productive. Knowledge of various types is a major component of human capital, and some taxonomy along the lines of generality or portability of the information is useful here because knowledge of different types will be paid for by different agents. The most general knowledge currently is taught in schools, but that is a relatively recent phenomenon, involving basic literacy, numeracy, possibly some introductory scientific facts and principles, possibly some information on one's society or culture (for example, history),

possibly some arts. This characterization may seem especially time-bound, but it provides a useful benchmark for what people at different times and in different cultures have known – and where they got the information. The general run of the populations in the ancient Mediterranean Basin were not widely literate (Harris 1989, esp. Ch. 1), although basic numeracy probably was more common – the demand for literacy was not high and its costs probably were high, while fingers and toes were general issue and measurement and counting were valuable in everyday activities. There is some information on scribal schools, but little on how the general run of the population gained whatever extent of numeracy they possessed. Of the four arithmetic operations, addition and subtraction are more intuitive and amenable to hand-and-foot technology, while formal schools, and even research, may have been required for multiplication and particularly division. The ancient version of societal or cultural knowledge, presently supplied by combinations of history, literature and social sciences, may have largely consisted of family and village tales and broader-based legends and myths, intertwined with religious education. Again, for the general run of the population, the family likely would have supplied the training. Such is the correspondence between contemporary and ancient in what we refer to here as the most general level of knowledge.

While general education does enhance productivity, and even the ability to gain occupation-specific knowledge, quite a bit of knowledge in any society is directed toward producing things. The generality of production knowledge lies along a spectrum with occupation-specific knowledge (itself sometimes called “general knowledge” in the human capital literature, but not referring to the general “schooling” knowledge we described above) at one end and firm-specific knowledge at the other. A person can take occupation-specific knowledge, such as carpentry or masonry, with him wherever he goes – between employers (firms), between locations. Firm-specific knowledge is of little value outside the firm where it is developed. It includes knowledge of specific routines that a producer may use. An example

of ancient firm-specific knowledge might be a farmer's detailed understanding of the properties and productive behavior of his land – knowledge that takes years of observation to develop, and may be passed on to one's family co-workers and heirs because they work the land day in and day out themselves, but would not be particularly transferrable to another owner of the land if the land were sold, simply because of the length of time it would take to convey the subtleties and multiple contingencies.

Location- and culture-specific human capital includes a variety of information from body language to who to bribe for an export permit. Language is an important body of such knowledge, which becomes more prominent when a large body of foreign-speaking immigrants is present in a society. The immigrants generally will find knowledge of the host language directly productive: they can “make more money” or keep from getting cheated as much if they understand the language. Among contemporary immigrants of similar backgrounds, even differences in facility with the host society's language can make a substantial difference in income.

Health and physical condition, the latter consisting of stamina and muscular strength, are also important elements of human capital. The value of physical condition in ancient societies probably was higher relative to the values of many forms of the available knowledge than is generally the case today because of the prominence of human muscle power among the mechanical energy sources in much of the technology of antiquity.

Saller (2012) uses concepts from human capital theory to excellent effect, accounting for such diverse phenomena as the typical age of apprenticeship to assessment of the price of a female slave in Egypt to the effect of Roman army service on agricultural productivity. Without appealing explicitly to human capital theory, Cohen (2002, 100) contends that “the Athenian concept of masculinity . . . – by relegating household operation and ‘slavish’ business pursuits to women – deprived Athenian men of economic opportunity and business experience” – clearly an instance of foregone on-the-job training.

10.2.1 *Investment in human capital*

As with most products, how much people invest in human capital depends on how much it costs to acquire and how much benefit they get from having it. It is natural to think of the cost of acquiring training (or education) as composed of tuition (payment to, say, an extispicist) and books (the livers?), but the largest component of training costs generally is the earnings the student foregoes while learning instead of working. At one end of the spectrum of foregone earnings costs is all training and no work, meaning that all earnings are foregone. The other end of the spectrum is not so clear-cut; it is possible to combine training and work – through either on-the-job training (ojt) or working part time and studying part time – but ojt reduces the earnings of a worker-learner (otherwise the training would be a free good). We will return to ojt shortly.

Investment in human capital frequently occurs over a lengthy period, which can pose problems in financing it. Without slavery, or at least legally enforceable indenture contracts, people find it difficult to borrow on the collateral of future labor earnings. Ironically, slaves very well might be better trained than their nonslave counterparts of equivalent aptitude because their owners are their financiers and can obtain better guarantees that their investments will be repaid – subject, of course, to the chances of mortality.³

On the benefit side of the issue, training must increase the productive capacity of the individual to have value as an investment. The more specific the training, the more straightforward it is to identify demands for it, but it can be more difficult to locate the demands for more general training. In fact, people can be not quite sure what skills they have to offer, which is one of the complicating elements of the labor market. Further clouding the training-productivity relationship is the possibility that training or education may offer a signal rather than raise productivity. According to this mechanism, signaling may not be directly productive but may be a reasonably reliable indicator of people who are productive. The ability to complete some type of training may let people reveal the extents of their intelligence, diligence, and perseverance – traits

which otherwise could take employers time and effort to determine.

Who actually pays for training depends on the type of training. General (occupation-specific) training will be paid for by the trainee, even if it is provided by an employer in the form of ojt. If any particular employer paid for the general training, the employee / trainee, once trained, could take the training to another employer, thus depriving the employer providing the training of the opportunity to repay himself in the form of the trained worker's higher productivity. General training furnished by an employer as ojt will be paid for by depressed wages during the period of training – but not necessarily a wage below the trainee's marginal product, net of the employer's training costs – as well as possibly a post-training period of wages below marginal product to permit the employer / trainer to recover the training outlays. The costs of ojt to an employer are multiple: reduced productivity of people assigned to teach; wasted materials and defective products; lower productivity of capital equipment before trainees learn how to use it properly; damages to equipment.

Firm-specific training is a captive set of skills, in the sense that pure firm-specific human capital would have no value to any other employer. Trainees' wages need not be depressed by the provision of firm-specific training. Most ojt provides human capital that falls somewhere on a continuum between purely general human capital and purely specific, with employers and employees sharing the costs.

Figure 10.1 shows the problem people solve in their choice of investment in human capital. The horizontal axis depicts the age of the worker from the time he or she can first either enter the work force or begin a period of training. Line E_1 is the time path of earnings if a worker chooses to acquire training. E_2 depicts the earnings growth the same person could expect to experience by foregoing training and immediately entering the labor force. If training is chosen, during the period of training (from time 0 to time a) earnings are actually negative as drawn, representing out-of-pocket costs such as tuition. At time a, the trained worker would enter the labor force and begin earning an amount per period of e_1^1 , which

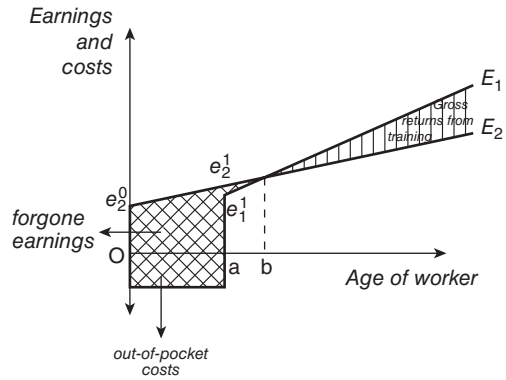


Figure 10.1 Costs and benefits of acquiring training.

is more than she would have earned as a brand new untrained worker at time 0 (e_2^0), but less than she would be earning at time a if she had started working at time 0 as an untrained worker (e_1^1) and experienced the growth rate of earnings (because of her growth in productivity deriving from experience). Having received training however, the graduated worker's earnings grow at a faster rate than they do for untrained workers, and at time b, her earnings overtake what they would have been had she not acquired the training. Thereafter, the differential earnings repay the sum of the out-of-pocket costs and the foregone earnings. At time 0, people deciding whether to acquire training or not contemplate these alternative earnings streams. The present discounted value of the gross returns (the vertically hatched area between earnings lines E_1 and E_2 after time b) must be equal to or greater than the sum of the two obliquely hatched areas of out-of-pocket costs and foregone earnings. The set of out-of-pocket costs, duration of training, initial trained earnings and the growth rate of trained earnings are not variables to be adjusted by the person considering training; they are part and parcel of the alternative occupation represented by earnings stream E_1 . Thus, the two earnings streams represent a choice of occupations as well as a choice of human capital acquisition.

On the basis of the earnings streams in Figure 10.1 we can calculate a rate of return to investment in human capital. That rate of return

should be comparable to the rates obtainable elsewhere. Nonetheless, there are uncertainties in the earnings stream from human capital. Possibly the most obvious one is the chance of death prior to one's expected life span. In times when short expected (adult) life spans have prevailed, the variability in those life expectancies may be correspondingly high, both factors retarding investment in human capital. The other principal uncertainty is the ability to sell the services of the human capital investment. The less variety available in human capital specializations, the less of a problem this source of uncertainty should be.

Training may be provided in different institutional and contractual settings. In contemporary industrial societies, most training is provided by state and private schools, some by trade and craft organizations through apprenticeships, and a minuscule amount within the family. This set of proportions probably was roughly reversed in the ancient societies of the Mediterranean Basin. The prominence of the family in the provision of training may very well have been a consequence of weak markets for insurance and pensions (both types of capital markets – or stated alternatively, intertemporal claims markets). Combining training with adoption by childless Egyptian scribes suggests that two capital-market problems may have combined: the weakness of pension (retirement income) provision for an aging scribe and the difficulty an aspirant scribe may have had in securing a loan to pay for training.

10.2.2 *Health*

Health is a bodily condition that affects multiple dimensions of productivity, from agricultural field operations to fertility. Within the parameters governed by one's physical environment and level of health-related knowledge of society, health status is a produced good, or condition. Within some limits, it can be improved, principally with nutrition; it can be "run down" during periods of particular need or emergency such as agricultural harvest periods or war without permanent, adverse effect.

Health status can be produced with various combinations of nutrition, labor time, and some public goods such as sanitation systems and public water supplies. Greater health status ("more health") would be reflected in greater muscular strength and stamina, but not necessarily in a constant proportion across individuals. Different jobs require different physical inputs, and these tasks commonly are divided among population subgroups (for example, male / female; child / adult / prime-age / elderly) along the lines of these requirements. Families are the primary loci of decisions regarding the distribution of health among a large proportion of the population. Because the inputs that produce health – time, including household time, and nutrition from various foods – are in scarce supply, it is generally the case in low-income societies that there are tradeoffs among the health of various family members. Those who stand to bring in the greatest amount of output (income), and those who respond most sensitively, physiologically, to the available health inputs, are likely to be invested in relatively heavily within families, a topic we shall revisit in section 10.7.

As human capital inputs into production activities, the various dimensions of health and different kinds of knowledge could be substitutes or complements, depending on available technologies with which they are employed. Different activities will, of course, use them in different proportions – think about, say, clerical work and farming.

10.2.3 *Guilds, occupational licensing, and entry restriction*

A larger supply of participants in any particular field of endeavor will depress the earnings of current participants, which will in turn depress the value of the human capital investments those people have made. One means of protecting such human capital investments from encroachments is to give authority to a group of such people already having invested in a particular type of human capital to review who may be allowed to work in the occupations using that type of human capital. A contemporary example is the

American Medical Association, which has substantial influence over licensing of doctors and medical facilities. Such organizations invariably refer to their activities benignly as providing quality control and hence improving the well-being of customers, but their control over the supply of the services is inescapable. To not act in their self-interest – that is, taking advantage of their control over entry to raise their own incomes – would be unexpected.

We can call such an occupational organization based on the possession of particular levels of skills a guild. A union, also common in contemporary labor markets, is similarly engaged in entry restriction with the aim of raising the incomes of members at the expense of nonmembers. Unions also may be involved in providing training and establishing training requirements for members, but they operate in occupations over which entry control is less completely exclusionary. Nonunion workers may operate in the same markets (areas) as union workers, but generally at a lower wage.

10.3 Labor Supply

10.3.1 *Utility analysis of individual and family labor supply*

Decisions about work are decisions about how to use time, so it will be useful to begin our consideration of labor supply with a simple accounting system for time. The simplest time accounting system divides total time available (T : the 24 hours in the day) into leisure (L) and work hours (H): $T = H + L$. In our first rounds at analysis we will use this classification. When thinking in particular about societies in which much of people's working time is spent working within the household, it is useful to refine the accounting system to distinguish between time spent working outside the household (H_m) from time spent working in it (H_h) and between household working time and true leisure: $T = H_m + H_h + L$. When our goal is to study decisions about how many hours to spend working outside the household, little or nothing is gained by disaggregating

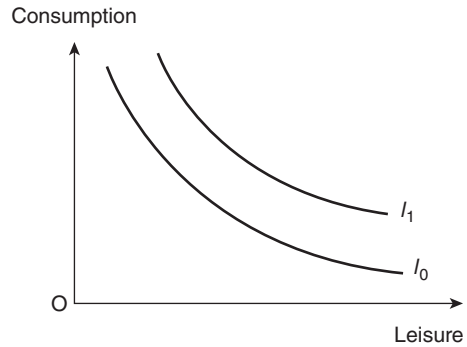


Figure 10.2 Indifference curve between consumption and leisure.

home time into leisure and household working time, and accordingly we will dispense with that distinction.

In this simple dichotomy of time allocation, the time not spent in leisure is spent working. The virtue of this observation is that, by treating leisure as a good that people value and demand like any other consumption good, we can study labor supply with the well understood tools of demand analysis. The demand for leisure implies the supply of labor. We develop a diagrammatic exposition of the demand for leisure. Figure 10.2 shows the indifference curve between leisure and consumption, and Figure 10.3 develops the earnings opportunities outside the household. The indifference curves between leisure and consumption (not of any good in particular but of all goods as a composite), represented by I_0

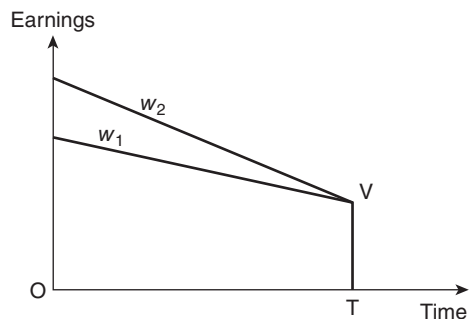


Figure 10.3 Earnings opportunities.

and I_1 , indicate the same kind of tradeoff we find between any two ordinary commodities: if some leisure is removed from a consumer, he must be compensated with some additional consumption if he is to be kept at the same level of satisfaction. I_1 represents a greater level of satisfaction than I_0 , and the fact that it is further from the origin of the graph than I_0 indicates that leisure is a “normal” good, in the sense that when a person’s income increases, as accounts for the difference in position between I_0 and I_1 , he will want to consume more of it.⁴ The slope of these indifference curves at each point on them is the (negative of) the ratio of the marginal utilities of consumption and leisure. This diagram represents the consumer’s preferences, or tastes, regarding the consumption of goods and leisure. Tastes are an important determinant of labor supply, but so are costs, which we introduce in Figure 10.3.

The axes of Figure 10.3 contain some subtle changes in nomenclature from those of Figure 10.2, but they measure equivalent concepts. Time, rather than leisure, occupies the horizontal axis – leisure being one possible use of time – and earnings are measured on the vertical axis – in real, as contrasted to monetary terms so that we can measure earnings in terms of real consumption. We need not extend the horizontal axis out indefinitely to the right, because we need account for only 24 hours, represented by T . The vertical line at T , extending up as far as V , represents the earnings this consumer has when she devotes her full amount of time to leisure; at point T , $T = L$. Clearly, if this amount of income requires no working time, it is nonlabor income; we can call it asset income. A longer vertical line would represent greater asset income. The two lines extending upward and to the left from point V look like the usual budget, or relative price, lines we used in consumption diagrams in Chapter 3. They are. As consumers reduce the amount of leisure they consume, we move from point T on the horizontal axis back toward the origin, and the slope of either of the relative price lines w_1 or w_2 indicates how much more earnings she can get by shifting time from leisure to work. These slopes are wage rates, w_2 representing a higher wage rate than w_1 . You can anticipate

that the next step with these diagrams is to overlay one upon the other and look for tangencies between leisure-consumption indifference curves and these budget-wage lines. Before proceeding with that kind of analysis, we turn to the logical structure of the consumer-labor supplier’s decision.

Treating leisure as an object of consumption and aggregating lots of different goods and services into a single aggregate we call consumption, we can write a consumer’s utility function as depending on the quantities of the composite consumption good, C , and leisure, L , he consumes: $U = U(C, L)$. This function has the usual properties: utility increases when either consumption or leisure increases, and it does so at a decreasing rate. In this functional notation, $\Delta U/\Delta C > 0$, $\Delta U/\Delta L > 0$, and $\Delta(\Delta U/\Delta C)/\Delta C < 0$, $\Delta(\Delta U/\Delta L)/\Delta L < 0$. Additionally, $\Delta(\Delta U/\Delta L)/\Delta C > 0$ and $\Delta(\Delta U/\Delta C)/\Delta L > 0$, important “cross-effects” which say that, at a given level of leisure and corresponding marginal utility of leisure, giving the consumer some more consumption will raise the marginal utility of leisure, and correspondingly for the effect of raising leisure with fixed consumption. The consumer maximizes this utility function subject to two constraints, one on the amount of “monetary” resources available and one on the amount of time available. The time constraint is $H + L = T$, and the budget constraint is $pC = wH + V$, where V is the amount of nonlabor income available. We can rewrite the budget constraint as $C = (w/p)H + V/p$, which shows that the real wage, w/p , is the slope of the budget line in Figure 10.3. We also can see from this rearrangement that when the consumer does not work at all ($H = 0$), he can consume an amount of C equal in value to V/p , his asset income. If he devoted all his time to working and none at all to leisure, the quantity of the consumption good he can have is $(w/p)T + V/p$, which is the measure of “full income” used in the household production model introduced in Chapter 3. If we substitute the time constraint into the budget constraint, we get $wT + V = wL + pC$, which offers some interesting insights. The left-hand side is full income in nominal terms (i.e., not

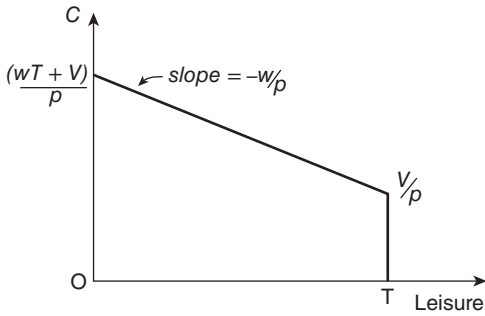


Figure 10.4 The budget line relevant to labor supply.

normalized in terms of relative prices by dividing the wage and nonlabor income by the price of the consumption good). The first term on the right-hand side is the consumer's "expenditure" on leisure – showing that he has to "purchase" leisure by removing time from work, the cost of doing which is the wage rate. The cost of leisure is the wage rate. The second term on the right-hand side is his expenditure on the composite consumption good.

Figure 10.4 shows this joining of the time and budget constraints, with the resulting information that (i) the slope of the budget line is the relative price of labor in terms of the consumption good, $-w/p$, (ii) the consumption available with no work is V/p , and that the consumption available if no leisure is chosen is $(wT + V)/p$, the vertical intercept of the budget line. The first-order conditions from the consumer's problem of choosing a combination of consumption C and leisure L to maximize his utility subject to the full-income constraint are $p = \Delta U/\Delta C$ and $w = \Delta U/\Delta L$, which tell us that he chooses an amount of consumption that equalizes the utility he gets from the last unit of it to the cost of acquiring it, and similarly with his "purchase" of leisure. Then the relative price of leisure and consumption (the "real" wage) is $w/p = (\Delta U/\Delta L)/(\Delta U/\Delta C)$. Recognizing that changes in the amount of leisure consumed are equal and opposite in sign to changes in labor supplied, we have the relationship that $\Delta U/\Delta H = -\Delta U/\Delta L < 0$, so that the real wage is also equal to the negative of the ratio of the marginal disutility of working to the marginal utility of consumption:

$w/p = -(\Delta U/\Delta H)/(\Delta U/\Delta C)$. Recall also that the right-hand side of this relationship is the slope of one of the leisure-consumption indifference curves at any point. Thus, the tangency of one of these indifference curves to one of these w/p wage-budget lines represents a choice of labor supply that maximizes the consumer's utility from consumption and leisure.

Now we are ready to show the determination of hours worked, and how changes in asset income and the wage rate affect that choice. Since an option is to not work at all, the choice of hours worked is a joint decision of what is called "participation" (participation in the labor market) and work hours. The decision to participate or not is determined by the consumer's "reservation wage," or the lowest wage that will induce her to forego any leisure (alternatively, the highest wage at which she *won't* work). At a wage just below this level, leisure seems very cheap to the consumer, relative to the price of consumption, given her property (asset, nonlabor) income. The utility she would lose by moving some time from leisure to work, and thence transform the potential productivity of that leisure into purchased goods, is large relative to what would be gained thereby. Thus, the additional consumption goods obtainable from supplying labor are not worth their cost in foregone leisure. We find the reservation wage by looking at the indifference curve that goes through point V/p ; the slope of the indifference curve at that point is the reservation wage, as shown in Figure 10.5, where w_r/p is the reservation wage.

Continuing on Figure 10.5, if the available wage were w_1/p , this individual would be indifferent between not working at all or working $T - H_0$ hours, measuring backwards from T . However, since the indifference curve I_0 cuts the budget line $-w_1/p$ at H_0 , the relative cost of leisure (w_1/p) does not equal the ratio of marginal utilities at either H_0 or T . Consequently, the consumer could reach a higher level of self-evaluated wellbeing on a higher indifference curve, I_1 as drawn, by supplying fewer hours at H_1 (hours worked being $T - H_1 < T - H_0$). As the wage continues to rise (that is, as we continue to make it rise in our analysis), the person will then increase the hours worked – up to a point, which leads us to

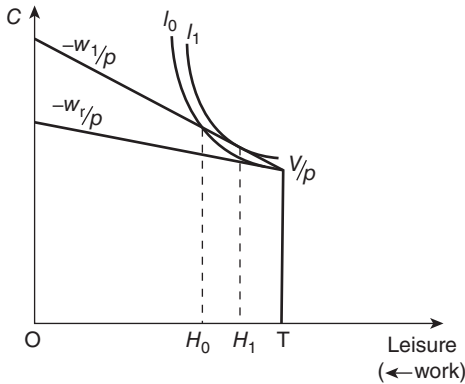


Figure 10.5 Influence of the wage on labor force participation.

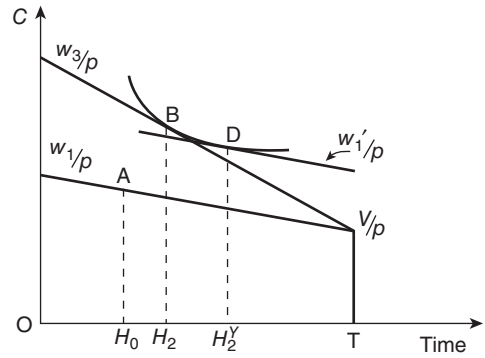


Figure 10.7 Income and substitution effects in labor supply.

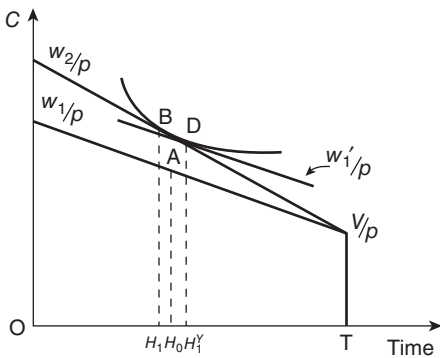


Figure 10.6 Influence of the wage on labor supply.

consider the substitution and income effects on labor supply.

In Figure 10.6 we trace out the effect that a higher wage has on a consumer's hours of work. Initially the wage rate is w_1/p , and, although we do not show the indifference curve tangent to that budget line, its tangency is at point A, where the consumer works $T - H_0$ hours. The wage rate pivots around point V/p from w_1/p to w_2/p . The tangency of a member of this family of indifference curves with budget line w_2/p occurs at point B, where the consumer supplies $T - H_1$ hours to work, which is more than $T - H_0$ hours. The pure income effect of this increase in the wage rate is represented by moving a budget line parallel to w_1/p outward to a tangency with the

indifference curve that is tangent to the w_2/p budget line (line w_1'/p). This tangency occurs at D, which would result in the consumer choosing to work H_1^Y hours (the Y superscript indicates that this is the work-hours choice resulting purely from the income effect). The movement along the indifference curve from D to B is the substitution effect. The substitution effect always draws more hours out of leisure and into work when the wage rate rises. To understand why the income effect works as it does, recall that leisure is a normal good: a higher income would induce a consumer to purchase more of it. Reversing this relationship for hours worked, we find that the income effect on hours worked is always negative. The net result can be either positive or negative; in fact, it is likely to be initially positive but ultimately negative, resulting in a backward-bending supply curve of labor for an individual. Figure 10.7 shows the wage rate increase even further, to w_3/p , and indeed the hours worked at w_3/p are fewer than those worked at w_1/p . Figure 10.8 plots hours worked against the real wage as a backward-bending supply curve, S_H .

Empirically, the magnitudes of individual labor supply elasticities estimated for contemporary societies vary considerably between adult men and adult women. For men, most estimates are small, ranging from 0.00 to 0.10, consisting respectively of substitution elasticities of 0.10 and 0.50 and income elasticities of -0.10 and

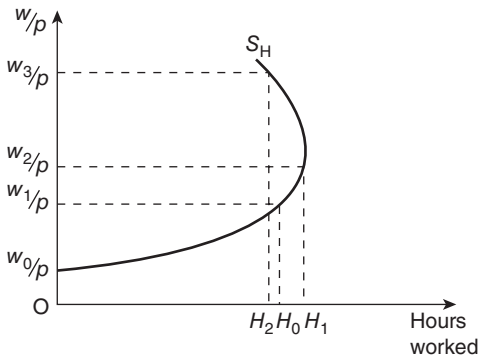


Figure 10.8 A backward-bending labor supply curve.

−0.40. Adult women, whose labor force participation rates are still considerably lower than men's, despite considerable increases in recent decades, have more elastic work responses to wage changes, with total, individual supply elasticities ranging from 0.45 to 1.35, composed of substitution effects of 0.50 to 1.65 and income elasticities of −0.05 to −0.30; a middle-range estimate for women is 0.80, with a substitution elasticity of 1.00 and an income effect of −0.20 (Killingsworth 1983, 193–199, Table 4.3). We will see in section 10.3.3 that aggregated supply elasticities are considerably larger. These adjustments are not instantaneous. When a wage change occurs, it can take from 2½ to 4 years for an individual to make half of the adjustment in his actual labor supply toward the desired level. Many aspects of one's life – daily, weekly, and longer time use patterns – are bound together and are costly to alter. The elasticities reported here describe the full-adjustment, or long-run, response. The extent to which these contemporary magnitudes offer useful guides to what their ancient counterparts might have been is an important question without a straightforward answer. Inasmuch as these magnitudes, and the male-female differentials in them, are produced by circumstances that may have characterized ancient labor milieus, they may not be bad guides or at least benchmarks. We will look at a couple of situations in subsection 10.3.4, on household production, that may help us

understand how to make informed inferences on the ancient magnitudes on the basis of contemporary evidence.

We should observe the differences here between transitory and permanent wage changes. The transitory wage change might be both seasonal and predictable, such as the higher wages available during peak agricultural seasons, but it is known – or at least believed initially – to be a temporary phenomenon. Transitory wage increases – or decreases – will elicit substitution responses in labor supply but not income effects. Thinking back to the permanent income theory of consumption and saving, differential transitory income will tend to be mostly saved instead of consumed. Permanent wage changes will produce both substitution and income effects in labor supply responses. One of the reasons for the slowness of response to wage changes, leading to the long lags between partial and full adjustment of the short- and long-run labor supply curves noted just above, is the difficulty people may have in distinguishing between transitory and permanent wage movements – and between price-level movements and relative price changes. We will see this distinction between effects of transitory and permanent wage changes from a slightly different perspective in the lifecycle treatment of labor supply in the following subsection.

As some contemplation of the income effect on labor supply might lead one to suspect, a larger nonlabor income raises the reservation wage, requiring a larger wage to pull a consumer into offering any time for work. Figure 10.9 illustrates this effect with a series of four indifference curves from the same family of indifference curves. Reading up the nonlabor income at time T (all time spent in leisure), the asset incomes increase as we go from V_0/p up to V_3/p . Drawing the real wage rates tangent to the lowest and highest of these four indifference curves, it is clear that the reservation wage at the highest asset income level is higher than the reservation wage at the lowest asset income level. This result is not an artifact of our drawing but a consequence of the positive cross-effects of consumption and leisure on the marginal utility of the other, which we noted above. The result, of course, is to require a higher

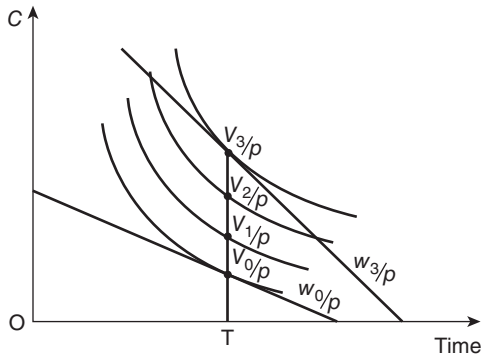


Figure 10.9 Effect of nonlabor income on labor supply.

wage to draw wealthier consumers out of their leisure and into work.

Although we are not using the household production model right now to analyze labor supply decisions, this simpler model carries the implication that leisure and purchased goods are consumed together. From the utility function in consumption goods and leisure, which our consumer maximizes subject to the budget and time constraints, we get a demand function for leisure, just as we got demand functions for goods in Chapter 3. Of course, we also get a demand function for our composite consumption good from this model as well, but we could actually enter as many different consumption goods as we wanted. Our demand function for leisure, $L^d = L(w, p_i, Y)$, is a function of all goods prices as well as the price of leisure (the wage rate). We know that $\Delta L^d / \Delta p_i < 0$ with income held constant, but $\Delta L^d / \Delta p_i$ can be of either sign, depending on whether good i and leisure are substitutes or complements. Since the demand for leisure is equivalent to a supply of labor, the supply of labor depends on the prices of all goods. Pursuing this thinking, it is easy to see also that the demand function for each good is a function of the price of leisure (independently of income) as well as of goods prices, implying that the demand for goods depends on the price of leisure and the ability of the good to substitute for or complement the consumption of leisure. The household production model lets one analyze these relationships between various purchased

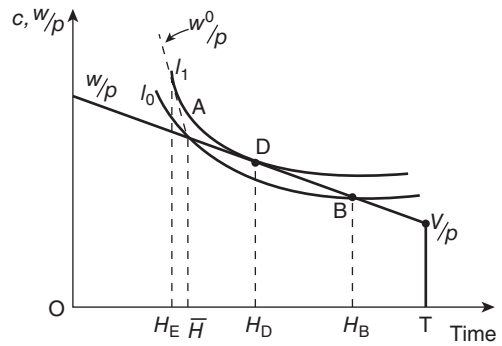


Figure 10.10 Fixed hours of work.

commodities and time spent in the household in greater detail, but the essential consumption relationship between commodities and leisure exists already in the simpler model.

Let's consider several specific topics in labor supply: What is the effect on labor supply and work behavior of all-or-nothing work offers by employers? How do various costs of working, such as commuting and child-care costs, affect labor supply? Figure 10.10 sets up the problem of the all-or-nothing work offer. An employer (the only one available for the worker under consideration) gives workers the option of working either $\bar{H}$ (measured to the left from T) or zero hours at wage rate w/p . I_0 is the indifference curve that passes through $\bar{H}$ at the going wage rate (point A). Since this indifference curve cuts the budget-wage line at A from above (rather than from below as at point B), this worker's marginal valuation of leisure is greater than the wage rate when he is forced to work $\bar{H}$ hours, and he would have a tendency to shirk. Notice that, although the indifference curve through V/p is not drawn, if it were, it would be below I_0 , yielding lesser satisfaction than working either $\bar{H}$ or H_B hours, so this worker would accept the all-or-nothing offer and shirk. He would maximize his utility by working H_D hours at wage rate w/p , but that is not an option. If the employer were to require H hours of work at wage w/p and offer additional work beyond those hours at the higher, over-time wage rate w^0 , he could not expect to elicit the full amount of time H_E , but he could get more out of this worker than H_B (the work hours minus

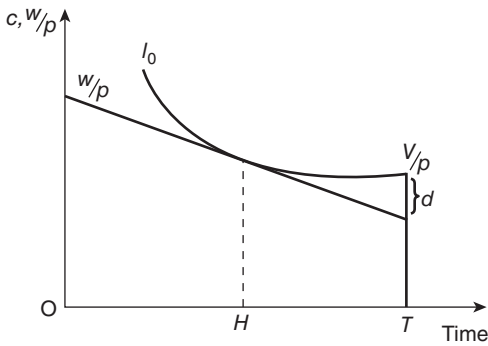


Figure 10.11 Effects of costs incurred to work on labor-force participation.

the hours shirked) – in fact H_D hours, with the difference between H_E (for which the employer pays) and H_D representing shirking.

We can study the effect of in-kind (out-of-pocket) costs of working, such as payments for child care, by reducing the height of the effective nonlabor income line, V/p . In Figure 10.11, a worker with nonlabor income V/p , who had to pay the amount d per work period, would be indifferent between working H hours and not at all. Let's take the same cost to a with-and-without comparison in Figure 10.12. Our worker/consumer begins with asset income V/p and wage w/p ; given this income and relative price, he would work H_0 hours (measuring leftward from T). Now, we confront him with an out-of-pocket cost of d per working period but

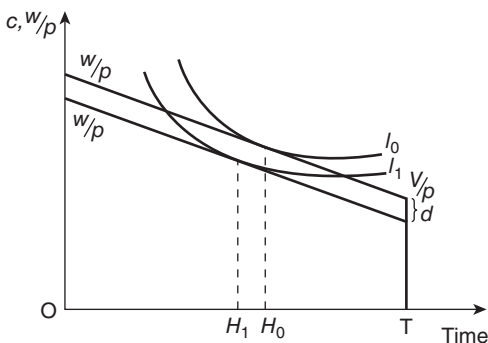


Figure 10.12 Effects of costs incurred to work on labor supply.

still let him earn the same wage when he works. We drop the budget line down parallel to the original one, but at a distance d below it. The highest level of indifference this worker can reach is now $I_1 < I_0$, and he would choose to work $H_1 > H_0$ hours. The general principle at work here we have seen already: this is a pure income effect on the demand for leisure.

Commuting costs – such as travel from one's village to an employer's fields in the case of a landless farm laborer – involve time, but not necessarily out-of-pocket costs. In this case, the fixed cost per work period is equivalent to a horizontal displacement of total time availability T , shown in Figure 10.13. Without the commuting time cost, with nonlabor income V/p , the worker would supply H_0 hours. After debiting the commuting time and starting the budget line from the vertical distance V/p above T' instead of above T , the work hours supplied fall to H_1 , measured leftward from T' (not from T). Both leisure and work hours fall – the commuting cost need not come fully out of either allocation – but the proportion of the reduction is an empirical matter. A cost that entailed both fixed time and fixed out-of-pocket costs would involve joint vertical and horizontal displacements of V/p and T .

Much of the same analytical structure carries over to the treatment of labor supply by family units, but the family does introduce some complications, the empirical importance of which may vary according to the topic one studies. We

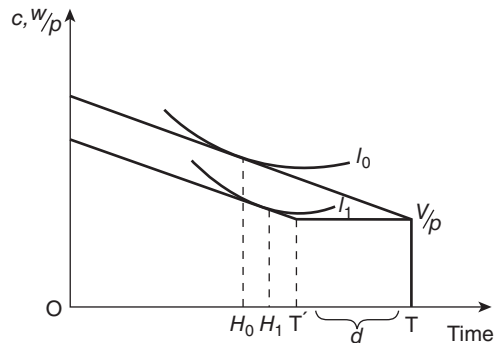


Figure 10.13 Effect of commuting costs on labor supply.

will open these complicating issues gradually. Initially, let's consider a two-earner household – a husband and a wife. From the perspective of the husband, the wife's earnings look like nonlabor income – he doesn't have to change his own time allocation to have access to that income. Similarly from the wife's perspective – her husband's labor income has the characteristics of asset income. Consequently, an increase in one spouse's wage rate affects both spouses' work choices, but for different reasons. The spouse whose wage rate changed faces a change in the cost of leisure but no change in asset income; the other spouse experiences an effective change in asset income, in the same direction as the change in the other spouse's wage, but no change in the cost of leisure. It is possible that one or neither spouse will increase their labor supply when a wage rate rises: the one facing the wage increase could decide to work less, and the other spouse definitely will work less because of the equivalent to an increase in asset income. Look at some of the previous diagrams to work through these changes.

What's wrong with – or oversimplified about – this picture? We can get some insights from setting up the family's utility maximization problem formally. Suppose there is a single family utility function based on the leisure and goods consumption of each family member, and a single, aggregate budget constraint into which each family member's time can be substituted. The n -member family's problem is: Maximize $U = U(L_1, L_2, \dots, L_n, C_1, C_2, \dots, C_n)$, subject to $\sum_i p_i C_i = \sum_i w_j H_j + V$, where H_j is the work hours supplied by family member j (with leisure L_j related to H_j by the total time constraint $T = L + H$). There will be substitution effects among the work hours of various family members that come purely through the utility function rather than through implicit asset-income effects. To justify the simple assessment of the problem we used in the previous paragraph, we need to assume such substitution effects are zero – empirically they may or may not be important, so there is not necessarily great damage done to predictions based on that assumption. This formulation of the family labor supply problem is called the individual

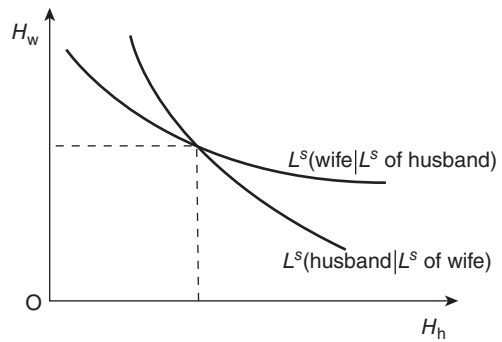


Figure 10.14 Reaction curves of husband's and wife's labor supply.

utility-family budget constraint model. This formulation of the utility function says that family members don't coordinate their labor supply. An important alternative formulation would let family members – say, the spouses – determine their hours of work contingent upon the hours of work of the other spouse, which formulates each spouse's work decision as a reaction function determined jointly with the other spouse's decision, as in Figure 10.14. In that figure, the subscripts w and h represent wife and husband, and the expression " L^s (wife | L^s of husband)" means "labor supply of the wife, given the labor supply of the husband." The intersection of the two reaction curves determines the equilibrium supply of work hours of each spouse.

Further complications of family labor supply involve the unitary specification of utility – a single utility function represents the entire family – and the determination of allocations within the household, but we delay treatment of those subjects until section 10.7.

10.3.2 Lifecycle / dynamic labor supply

In the analysis of labor supply in section 10.3.2, the consumer and potential worker looked at a single period – a day, week, month, year, or whatever division of time – and decided how much leisure to consume and how much to work. In the shorter periods – days to months – it seems reasonable that people would tie their

decisions about how much to work in one period to how much they plan, or are able, to work in nearby periods of the same length. In the longer periods – years – labor reveals its capital-like characteristics, and the major events of the human lifecycle – gaining strength in youth, developing a family and having children, losing strength in old age – impose predictable patterns of labor and leisure choices. The single-period, or “static,” analysis yields many useful insights into time allocation decisions and should not be discarded, but the lifecycle, or “dynamic,” analysis contributes further insights.

Lifecycle labor-supply analysis considers all the potential working periods of an individual's life, from his or her first “entry into the work force” (in some of our ancient Mediterranean societies, about the time a three- or four-year-old is big enough to carry some straw or stir a pot) until the anticipated time of death. The concept of anticipated time of death may give some students pause on the grounds that such a date is so uncertain. Granted, but people have a concept of a “normal” lifespan, around which they place a modern or ancient version of a probability distribution for themselves and base their planning for how they will arrange their activities across the major phases of the lives they can expect if they are lucky. Stochastic lifecycle labor supply analysis is distinctly more complicated than the complete certainty case we rely on here and doesn't really add that much we will find of value for the effort.

We begin with an individual's lifetime utility function, which contains her consumption and leisure choices in each period of her life: $U = U(C_0, L_0, C_1, L_1, \dots, C_T, L_T)$. She maximizes her utility subject to a series of wealth constraints for each period, plus a terminal-period constraint that says she can't go out owing to people (this last constraint is really a convenience for us as analysts, since it lets us separate generations). In each period, the wealth constraint says that the accumulation of wealth between that period and the previous one equals the interest accumulation on the previous period's stock of wealth, plus labor earnings, minus consumption expenditures: $Z_{t+1} - Z_t = rZ_t + w_t H_t - p_t C_t$,

where Z is wealth, r is the interest rate, or rate of return available on investments, and H is hours of work. The terminal-period non-negativity constraint is $Z_{T+1} \geq 0$, which says that at the first day of the period following the individual's death, her assets can't be negative: she can leave a bequest, either accidental or deliberate, but not debt. If we constrain the terminal condition to be zero, $Z_{T+1} = 0$, representing no bequest, and use the relationship that leisure equals total time in each period (defined to be 1) minus hours worked, $L_t = 1 - H_t$, we can rewrite the series of wealth constraints as a single constraint: $(1 + r)Z_0 + \sum_{t=0}^T (1 + r)^{-t} w_t = \sum_{t=0}^T (1 + r)^{-t} (w_t L_t + p_t C_t)$. In this formulation of the wealth constraint, the left-hand side represents “full wealth” along the lines of “full income” in the household production model: the wealth that a person could have at each period t if she devoted all time to work and none to leisure and “banked” all labor income instead of consuming part of it; the right-hand side shows her “purchases” of leisure and of consumption goods.

An interesting relationship between one's valuation of one's assets and one's earnings opportunities emerges from the formal analysis of the lifecycle consumption-labor supply problem. The Lagrangean representing this constrained maximization problem is $\mathcal{L} = \sum_{t=0}^T (1 + \rho)^{-t} U(C_t, L_t) + \lambda [Z_0 + \sum_{t=0}^T (1 + r)^{-t} (w_t H_t + p_y C_y)]$. The consumer discounts her future consumption at her personal rate of time preference, ρ , while future assets accumulate at the interest rate r ; the two rates may, but need not, have the same value. The consumer adjusts the values of three variables to perform this maximization, two of which are time streams, the other a scalar (that is, a single number, independent of time): C_t , H_t (or, equivalently, L_t), and the Lagrange multiplier λ , which is the shadow value the consumer attaches to her initial-period assets. This valuation is determined simultaneously with her consumption and labor supply decisions; even though the value of λ is not observable, it is an important determinant of major, observable quantities. If the consumer “solves” for λ and comes up with a value that is “too low,” she will behave as if she really were wealthier than she is and consequently work too

little and consume too much, ending up in debt by her terminal period, if not well before.

The first-order conditions for this utility maximization have a familiar form. First, the marginal benefit of consumption equals its marginal cost in all time periods: $\Delta U/\Delta C_t - \lambda_t p_t = 0$, where $\lambda_t = [(1+r)/(1+\rho)]^{-t} \lambda$, which is the current-period valuation of assets, the shadow value of initial assets divided by the ratio of the interest rate to the rate of time preference; if those two rates are the same, $\lambda_t = \lambda$ in each period. Consider what happens to consumption if the consumer assigns too low a value to λ : a low value of λ would encourage the consumer to consume more in order to drive the marginal utility of consumption into equality with λ_t^* times p_t , where the calculated value $\lambda_t^* < \lambda_t$ (λ_t being the “true,” or equilibrium, value). Second, the marginal valuation of leisure is equal to or greater than its marginal cost in each period: $\Delta U/\Delta L_t - \lambda_t w_t \geq 0$; in any period in which $\Delta U/\Delta L_t > \lambda_t w_t$, $H_t = 0$, that is, the consumer does not work. Again, if the consumer calculates λ below its equilibrium value, she will be induced to consume more leisure than would satisfy the equality in the first-order condition for the leisure-labor choice. Finally, the consumer must stay within the wealth constraint: $Z_0 + \sum_{t=0}^T (1+r)^{-t} (w_t H_t + p_t C_t) = 0$, which says that initial assets, plus the discounted lifetime difference between labor income and expenditure must be zero (because we assumed away bequests and terminal-period debt), with any transitory surpluses of income over expenditure being saved while transitory deficits of current expenditure over current income must be covered by previous saving.

The dynamic equilibrium path of labor supply (hours worked; leisure is its mirror image) is the net result of three forces responding to the exogenous time path of wage opportunities. Figure 10.15 shows an equilibrium time profile of hours worked responding to an exogenous time profile of the wage. First, the existence of savings separates consumption decisions from work decisions and lets the consumer transfer effort from periods of low earnings to periods of high earnings in a pure efficiency effect. This is also known as an intertemporal substitution

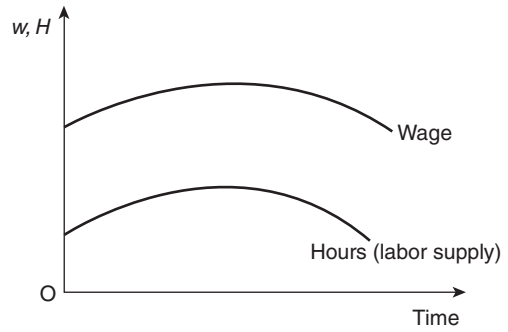


Figure 10.15 Labor supply and wage over the lifecycle.

effect, which is conceptually different from the atemporal, or single-period, substitution effect, the former being a movement along the time profile of labor supply and the latter being a shift in the labor supply curve. Second, the ability to invest earnings induces her to work more earlier in life, invest the earnings, and reduce her labor supply when she is older, assuming that leisure is a normal good. Third, a positive rate of time preference cuts against the interest-rate effect inasmuch as the onerousness of future effort is discounted relative to that of current effort (homework now compared to homework after a movie for a teenager), leading the consumer to prefer leisure earlier to later. If $\rho = r$ (the subjective rate of time preference equals the interest rate), hours worked and wages always move in the same direction over the lifecycle. If $\rho < r$, hours and wages may move in opposite directions at least over a portion of the lifecycle: if the wage rises rapidly early in the lifecycle but rises more slowly later, w and H will both rise early, but H will peak and begin to recede in the mid- to late lifecycle while the wage continues to rise. The time path of hours in Figure 10.15 is not a free-hand drawing, but the sum of these three forces at each point in time.⁵

Let's consider how the time profiles of hours (leisure) and consumption respond to changes in the time profiles of the wage and the consumption price and to the initial level of assets. We can interpret this comparison in two ways: either as how differences across individuals in

A_0 , w_t and p_t lead to differences in their H_t or as how changes in a particular individual's A_0 , w_t , and p_t would lead her to alter her labor supply time profile. Responses to anticipated changes in the wage correspond to movements along the labor supply time profile. These kinds of wage changes involve only substitution effects; the income effects (actually, the wealth effects) are already incorporated in the equilibrium profile of work hours supplied over the lifecycle. Responses to unanticipated changes in the wage involve shifts in the time profile of hours worked and can be analyzed in terms of the customary income (or wealth) and substitution effects. The unanticipated wage change can involve a shifting of the entire lifetime profile of wages, a change over some group of periods, or a change in only a single period. All three types of change can produce responses in periods outside the time during which the wage is different from what it was expected to be. Even if we rule out direct substitutability and complementarity of leisure and consumption in different periods, a change in w_t in even a single period will cause a revaluation of λ in the opposite direction to the change in w_t , and the revaluation of λ in turn will cause a shift in the entire time profile of hours worked (and of consumption as well).

Figure 10.16 shows such a shift in a time-profile of labor. The solid lines represent base equilibrium paths of H_t and w_t : our consumer has thought

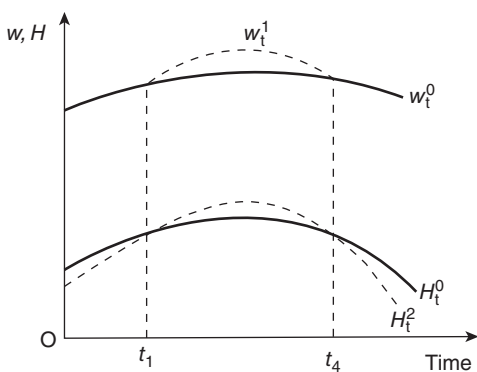


Figure 10.16 Responses to anticipated and unanticipated wage changes over the lifecycle. Adapted from Killingsworth 1983, 226, Figure 5.2. © 1983 Cambridge University Press.

about what her time path of wages is going to be and has decided on a time path of hours of work she'll supply at various times of her life. We can think of the H line in Figure 10.16 (and Figure 10.15 as well) as the percentage of available time devoted to work, calculated from $H = T - L$ and letting T equal 1). Next, suppose that between times t_1 and t_4 , the wage is expected to be higher than it is in the "base forecast." Granting the rudimentary character of long-term forecasting in the ancient Mediterranean lands, what could be the empirical correspondence to this hypothetical displacement of the wage over a future time span? Remember – the changes in the wage represented by the dashed lines are future events anticipated at the beginning of our consumer's working life. One interpretation of our wage displacement is that we are comparing the expected wage profiles of different individuals – we, as analysts, have the comparative knowledge while the actual consumers don't but just behave differently. Another interpretation is that an individual may be contemplating entering two alternative occupations (but the diagram ignores the differential training costs, which weakens but does not entirely vitiate the interpretation); possibly a young man is contemplating being a mercenary during this time.⁶ Under either interpretation, the higher wage from t_1 to t_4 affects labor supply over the entire lifecycle. If we assume for the moment that $p = r$, between t_1 and t_4 , the difference in hours worked is the net result of the efficiency effect of the wage raising hours and the indirect effect of a reduced shadow valuation of initial wealth lowering hours (λ : $\Delta\lambda/\Delta w_t < 0$). From 0 to t_1 and after t_4 the efficiency effect of the higher wage does not operate, the wage profile-induced reduction in λ being the sole influence on differential hours worked. As the figure is drawn, the wage differential does not raise work hours over their base supply immediately at time t_1 , but only at t_2 , when the direct wage effect outweighs the wealth effect of the lower λ associated with the higher wage profile during period $t_1 t_4$. Also as drawn, the wealth effect outweighs the efficiency effect of the higher wage before the higher wage period is over, with the labor supplied with profile H_t^1

slipping below the hours offered for the base wage profile at t_3 .

The general structure of response of the lifetime profile of hours worked to a change in the wage during some period can be decomposed into two basic elements: the direct effect of the wage differential during the period when it is effective and the indirect wealth effect. We can write these as $\Delta H_t / \Delta w_{t^*} = \Delta H_t / \Delta w_{t^*} |_{\bar{\lambda}} + (\Delta H_t / \Delta \lambda)(\Delta \lambda / \Delta w_{t^*})$, in which the subscript t on H indicates that we are referring to the entire lifetime profile of hours worked, and the subscript t^* on w indicates that the wage displacement refers only to the periods designated by the asterisk. The first term on the right-hand side is the direct effect of the wage change on hours worked, with the shadow valuation of initial wealth held constant. It operates only during the periods when there is a displacement of the wage relative to the reference wage; otherwise it is zero. The second term, working backwards from the term in the right-most parentheses, is the effect of the period of displaced wages on the shadow value of initial wealth, times the effect of the induced change in the shadow value of initial wealth on hours. This effect operates in all periods of the lifecycle, regardless when the period of displaced wages occurs.⁷ As we've noted, $\Delta \lambda / \Delta w_{t^*} < 0$. The other two major parameters influencing the shadow valuation of initial wealth are the initial level of assets and the consumption price: $\Delta \lambda / \Delta Z_0 < 0$ also, but the effect of p_t can be positive or negative, depending on the magnitude of consumption during the periods when the price displacement occurs and a measure of the substitutability between consumption and leisure; the greater is the level of consumption and the lower the substitutability, the more likely is $\Delta \lambda / \Delta p_{t^*}$ to be positive.

A modest difference in the lifetime wage profile can be expected to have a correspondingly modest effect on the lifetime labor supply profile, as we saw in Figure 10.16. A substantial difference in the wage profile will produce a large change in the labor supply profile, as shown in Figure 10.17, in which the shift from profile w_t^0 to w_t^1 precipitates the change from labor supply profile H_t^0 to H_t^1 .

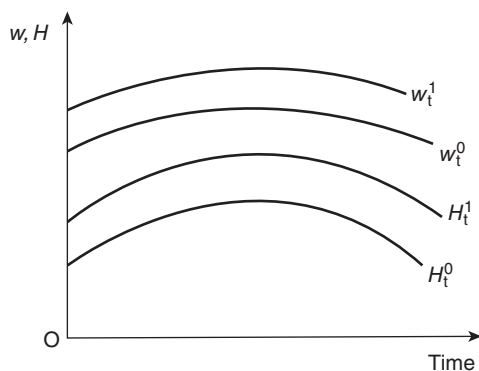


Figure 10.17 Effect of change in anticipated lifecycle wage profile on anticipated lifecycle labor supply.

10.3.3 Supply of labor to activities

So far we've considered the labor supply decisions of individuals, and we've seen how an individual's labor supply curve can be constructed by tracing out his hours worked at various wage rates. The individual responses are essential for understanding, but the individual is not the only perspective on labor supply that is useful. An employer in a single firm might wonder how the number of applicants might respond to a higher wage that she might offer. Similarly, an army's chief adjutant might wonder how many more recruits he might get if he were to raise the daily wage or ration. If we were to ask a similar question about, say, Egyptian scribes in the Middle Kingdom, one reaction might be that the number of scribes needed was simply determined by the palace, and their ration (wage or salary by another name) was assigned by custom, so where's the room for a supply curve? Possibly. But let's take a closer look at some events that might have escaped our notice. A family has the choice of having one of its sons become a scribe and others remain managing the family estate. Suppose that the prosperity of the estate rises – farm prices have risen, let's say. The patriarch thinks about which of his sons he wants to manage the prospering estate and which to become a royal scribe. The son who is a royal scribe will be in a position to deliver certain benefits to the family, so the selection of the designee for that occupation needs

care. However, the farm is really doing well, and to keep it running efficiently and prosperously needs an intelligent and aggressive mind. The “best” two sons are developed as estate managers, and poor Wen-Amun, who’s a bit slower but not altogether a dullard, can be spared for training in the prestigious royal occupation. The number of scribes altogether stays the same, as does their pay, but the occupation attracts less able entrants. The quality-corrected supply of scribes would certainly fall if all fathers have other opportunities for their sons.

Return to the theme of the degree of aggregation of a labor supply curve, or the market or other entity whose responses it describes. The smaller the entity looking at labor supply responses facing it, the more elastic is the supply. The individual firm, assuming that it is small relative to all demanders of labor in the local economy, sees a flat labor supply curve. As far as it is concerned, it can hire as many people as it wants without affecting the wage it has to pay to get them. This characteristic will be patently untrue for large firms that employ a substantial fraction of several of the skills – occupational groups – in the local economy. The labor supply curve for a particular occupation – say, clerks, or even something as broadly defined as “professionals” – will have considerable slope to it, but it still will be quite responsive to wage offerings. Figure 10.18 shows short-run and long-run supply curves of labor to a particular market or occupational group.

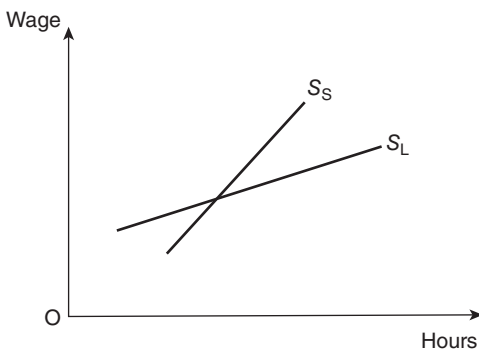


Figure 10.18 Short- and long-run labor supply curves.

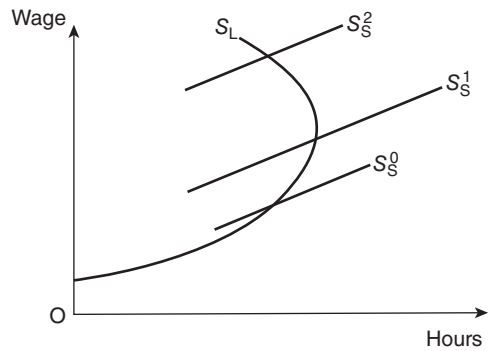


Figure 10.19 An economy-wide labor supply curve.

The short-run curve is considerably less elastic than the long-run curve because of the costs of switching occupations or entering one for the first time. Short-run elasticities for even broad occupational groupings in contemporary economies are commonly as high as 1.0, while long-run elasticities range as high as 3.0–4.0.

In contrast to market labor supply curves, the aggregate labor supply curve to an entire economy looks much more like the individual labor supply curve, as shown in Figure 10.19. When we consider the supply curve facing an individual market, the entries into the supply at the upper end of the curve come from other markets, but when we consider the supply facing an entire economy, virtually by definition there are fewer opportunities for entry from “outside” when the wage rises. Population growth occurs slowly, and adjustments to birth and mortality rates occur even more slowly. Immigration is one such source of entry, but except in exceptional cases, the percentage increase in an entire economy’s labor force from immigration will be small.

10.3.4 Household production

We introduced the household production model in Chapter 3 to get additional insights on the demand for purchased goods. Here we will give the general model some workouts with particular problems that bear on how people use time,

particularly outside markets, since nonmarket work was a particularly important component of time allocations in the Mediterranean region in antiquity. The central core of household production theory is that consumption requires more than just monetary expenditures – or nonmonetary expenditures in a non- or poorly monetized economy. Time must be expended in production, either in “the market” (working for others) or working at home (for oneself presumably), to produce or otherwise acquire goods; once the goods are acquired, you have to do something with them either before or in the act of consuming them. Think of eating a meal or reading a book. In the analyses of labor supply so far, we haven’t distinguished between household work and leisure when studying the decision to work for somebody else. In this subsection we relax the focus on working for somebody else, although that will remain in the background as a major option, and link that location of work to other locations of work and to other, nonwork activities.

The utility function of household production theory contains “Z-goods” as its arguments. These Z-goods are the final consumption of articles that are either purchased outside the household or made within the household, with the consumption taking time beyond what was required to either purchase or make it. This utility function is $U = U(Z_1, Z_2, \dots, Z_n)$, in which subscripts 1 through n represent different final consumptions: meals, cleanliness, children, reading, playing games, hunting, sleeping, and so on. The production function for any particular consumption good Z_i is $Z_i(T_{zi}, C_{1i}, C_{2i}, \dots, C_{ni})$, in which T_z is time in a home activity (production of some Z-good), and the further subscript i indicates that we refer to the time spent in producing the i^{th} Z-good, Z_i . The fact that the production function itself is also subscripted – $Z_i(\blacksquare)$ – means that each Z-good has its own production function – that is, it requires a different combination of time and purchased goods C_j per unit of Z_i . The purchased goods in the Z-good production function are the C_1, C_2 , and so on, through consumption good n ; the additional subscript i (for example, C_{1i}) again indicates that in this production

function we are referring to the amount of purchased good C_1 used in Z-good i . We can substitute the Z-good production functions into the utility function to get utility as a function of nonmarket time allocated to each final consumption and the market goods themselves: $U = U[Z_1(T_{z1}, C_{11}, \dots, C_{n1}), \dots, Z_m(T_{zm}, C_{1m}, \dots, C_{nm})]$. The consumer faces budget and time constraints: $\sum_i p_i C_i = wH + V$ and $T = H + \sum_i T_{zi} + L$, where V is nonlabor income, H is hours worked outside the household, and L is leisure. We can substitute the time constraint into the budget constraint to get the “full-income” constraint. The consumer maximizes the Z-good utility function subject to the full income constraint, the first-order conditions for which yield shadow prices of final consumption that are combinations of time cost (valued at the out-of-household wage rate, or if there is no out-of-household work, at some numeraire household production activity) and purchased good cost. Consumption goods will be distinguished by their time intensity, that is, by the amount of time required per unit of final consumption, as well as by their purchased-goods intensity.

The first problem we address with the household production model is how extra-household work opportunities affect the allocation of time within the household. Figure 10.20 begins the analysis with a household production function – the convex curve with its horizontal intercept labeled T – and a Z-good utility function labeled U_0 . The convexity of the production function represents decreasing marginal productivity of

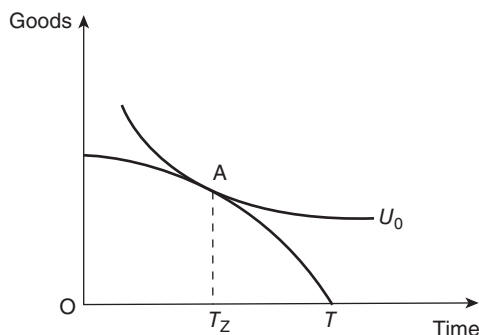


Figure 10.20 Household production function.

both purchased goods and household time. The horizontal intercept of the production function represents the total time available to the household as T . At the point of tangency between the production function and the utility function, labeled A, the household equalizes the ratio of marginal utilities of leisure and Z-good consumption (the diagram relies on a composite Z-good) and the marginal cost of transforming leisure into Z-goods through the use of household time to produce the Z-good. The household's time is divided between $T - T_z$ household working time and $T_z = 0$ leisure.

Figure 10.21 contains the same production function as Figure 10.20, with the point of tangency with U_0 indicated by point A. The straight line labeled w_0 represents the wage available by working outside the household. Its tangency with the household production function determines the allocation of time between household work and a combination of "market" work and leisure; $T - T_{z1}$ hours of household work are undertaken, a smaller number than the $T - T_{z0}$ worked in the household in the absence of the extra-household opportunity. At the allocation of $T - T_{z1}$ hours of household work and $T_{z1} - H$ hours of market work, the marginal productivities of labor in the household and outside the household are equalized. Indifference curve U_1 is a higher level of utility in the same family of indifference curves

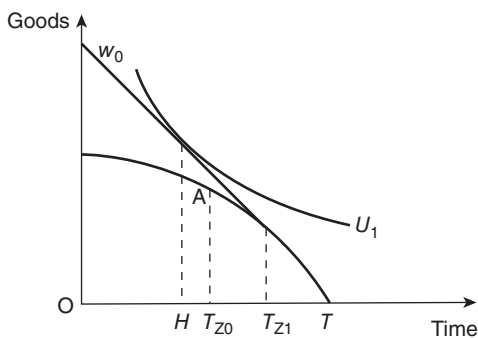


Figure 10.21 Household production frontier with out-of-household work opportunities. Adapted from Gronau 1986, 284, Figure 4.2. Reprinted with permission of Elsevier/North-Holland.

represented by U_0 in Figure 10.20, but represents a higher utility than was attainable with only household production. This is achieved with "market" work hours equal to $T_{z1} - H$, leaving $H = 0$ hours of leisure. At the combination of $T - H$ total hours working (at equal productivities at the margin) and $H = 0$ hours of leisure, the marginal valuation of leisure is equalized to its marginal cost. The introduction of the out-of-household work opportunity has resulted in a reallocation of work time from the home to the market, entailing an overall increase in work and a corresponding reduction in leisure. The reduction in leisure is relatively small, a result of large substitution and income effects, which we have not drawn but which you can either draw yourself or imagine.

Now, suppose that the wage available outside the household increases, as the shift from line w_0 to w_1 in Figure 10.22 represents. Household working time falls further, to $T - T_{z2}$, and "market" hours increase further, to $T_{z2} - H_2$ ($> T_{z1} - H_1$). Leisure also falls further, to $H_2 = 0$ ($< H_1 = 0$). Point A on the production function denotes the original, autarkic allocation of time between household work and leisure. The introduction of the outside opportunity and the increase in its attractiveness have affected household working hours more than they have leisure.

Returning to the autarkic situation depicted in Figure 10.20, Figure 10.23 gives some non-labor (asset) income to this household, in the

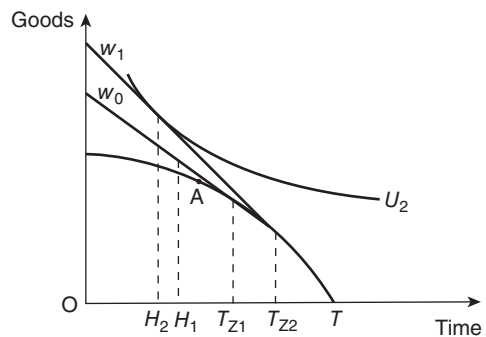


Figure 10.22 Effect of higher earnings on household production.

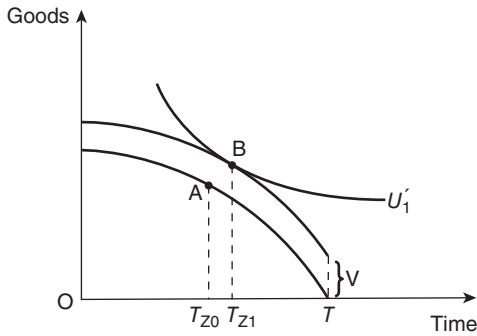


Figure 10.23 Household production with nonlabor income.

amount V . Point A in Figure 10.23 corresponds to point A in Figure 10.20, and indifference curve U'_1 (the “prime” indicating its distinctness from utility level U_1 of Figure 10.21) has its tangency to the vertically displaced production function at point B. Without any outside work opportunities, the increase in asset income induces the household to reduce its household working hours and increase its leisure. Compare the slopes of the production functions at points A and B, and you’ll see that the marginal cost of leisure is lower at B than at A.⁸

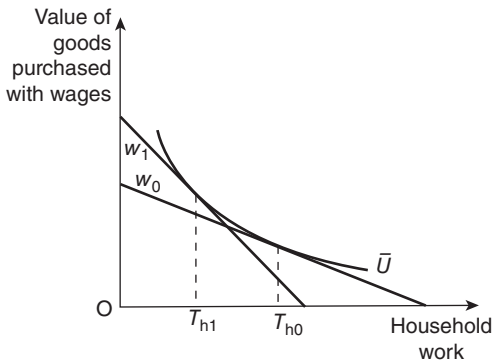


Figure 10.24 High substitutability between household- and market-produced goods. Adapted from Ronald G. Ehrenberg and Robert S. Smith, *Modern Labor Economics: Theory and Public Policy*, 7th edn, p. 229, Figure 7.2 (a), © 2003, 2000, 1997, 1994. Reprinted by permission of Pearson Education, Inc., Upper Saddle River NJ.

Figures 10.24 and 10.25, offer further insight into the sizeable differences in the responses of household work hours and leisure hours in the analysis of the “opening” of a “closed” household to outside employment opportunities.⁹ The key to the differential responses lies in the differential substitutability between household work and market work on one hand and between leisure and market work on the other. The gentle curvature of the utility function in Figure 10.24 shows easy substitution between household time and the goods that can be purchased with market time in the ability to deliver a given level of utility, U (think of leisure being on a third axis coming out of the origin in a three-dimensional diagram; similarly, think of Figure 10.25 with household work hours on a third axis through the origin). The same wage increase, from w_0 to w_1 , occurs in both diagrams. A large reduction in household work hours results, from T_{h0} to T_{h1} , with a corresponding increase in purchased goods used to keep utility at the same level; household and market work hours are good substitutes. Figure 10.25 depicts the utility relationship between leisure and market work and the same pair of wages, w_0 and w_1 . It is much more difficult to substitute the goods that market time can purchase for

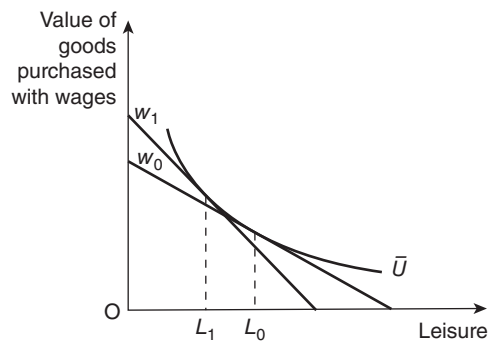


Figure 10.25 Low substitutability between household- and market-produced goods. Adapted from Ronald G. Ehrenberg and Robert S. Smith, *Modern Labor Economics: Theory and Public Policy*, 7th edn, p. 229, Figure 7.2 (b), © 2003, 2000, 1997, 1994. Reprinted by permission of Pearson Education, Inc., Upper Saddle River NJ.

leisure while keeping utility constant. Leisure can be reduced (replaced by market time) by only $L_0 - L_1$ and still keep utility constant.

The relative dominance of women in household production in recent times – say, putting them somewhere around T_{h0} until recent decades, may go a considerable way toward accounting for their much higher labor supply elasticities, and particularly the substitution components of those elasticities, reported in subsection 10.3.1. To the extent that contemporary women (at least in the industrial economies) have moved much closer to a household-market time allocation like T_{h1} in recent years, their labor supply elasticities – at least the substitution components of them – could be expected to fall in the near future. The shares of work time allocated to household and outside employment in ancient settings similarly would have been key to determining the sensitivity of those societies' labor supply responses to out-of-home wage offers. If both men and women worked predominantly for themselves in the home, the responsiveness of workers of both sexes to outside employment offers could have been considerable, even if not highly or directly observable in either textual or archaeological evidence.

We can develop somewhat deeper insight into different labor supply responses of men and women by exploiting the production function in the household production model.¹⁰ Consider a two-member family with a joint utility function $U = Z^c$, where the Z -good is produced by home time of both spouses: $Z = T_{hw}^a T_{hh}^b C^{1-a-b}$; T_{hw} is the household time (or leisure) of the wife, T_{hh} is the household time of the husband, the exponents a and b are coefficients of the constant-returns-to-scale Cobb–Douglas form of production function, and C is purchased goods. The exponent c on Z -goods in the utility function is equal to or greater than one. The household maximizes this utility function subject to the budget constraint $pC \leq V + \sum_i w_i H_i$ and the time constraint $H_i + L_i = T$ for each spouse (subscript i represents the two spouses). We rearrange the first-order conditions for home production time (or leisure) and find that $L_i = A_i F / w_i$, where $A_w = a$ for the wife and $A_h = b$ for the husband,

and the family's full income $F = V + T \sum_i w_i$. This expression for utility-maximizing home time reveals that even if the husband and wife earn the same wage outside the household, the wife will devote more time to the household, or nonmarket work, if $a > b$, that is, if her productivity in home production is higher than that of the husband. (Women readers will naturally recognize the gaming possibilities here.) A higher wage for the husband will reinforce this effect. We can pursue this reasoning to find a larger labor supply elasticity for the wife. Use the expression for L_i just given and the fact that labor supply is $H_i = T - L_i$ to get the expression for labor supply of each spouse. From that expression, some further rearrangements yield the elasticity of each spouse's labor supply with respect to his or her own wage: $\epsilon_{ii} = (A_i/H_i)[(F/w_i) - T]$. As long as $a > b$, $A_w/H_w > A_h/H_h$, which will make $\epsilon_{ww} > \epsilon_{hh}$, even if their wages are the same. If the husband's wage exceeds the wife's, the effect is reinforced. Thus, the differential home productivity yields the differential allocation of time between in- and out-of-household work that we took as a given in the previous example. Explanations for such a differential are not explored.

The final application of the household production model in this subsection demonstrates the capacity of this model to bring within the scope of normal economic analysis topics not ordinarily considered to be subjects of resource allocation decisions.¹¹ While there is undoubtedly a large, biologically determined component of the time people spend sleeping and considerable individual variation in "needed" sleeping time, survey evidence indicates 20–30% variation in daily sleeping time of individuals over the week and across seasons and years. Since roughly 25–30% of total time (the T we have been using in the models above) is devoted to sleep (including napping and sexual activity), it is potentially important to understand the scope for discretionary behavior regarding this large block of time. The restorative powers of sleep, its ability to sustain and improve productivity in a wide range of waking activities, reinforce this importance.

The household production model offers a natural vehicle with which to formulate an individual's demand for sleep. Suppose that sleep both contributes to productivity and is valued for itself. The wage outside the household has two components: $w = w_1 + w_2 T_s$, in which w_1 is a base wage, unaffected by the productivity contributions of sleep, w_2 is the component affected by sleep-induced productivity, and is time spent sleeping. The utility function contains Z-goods and time spent sleeping: $U = U(Z, T_s)$, $U_i > 0$, $U_{ii} < 0$, $U_{ij} > 0$.¹² We use a fixed-coefficients production function for the Z good for simplicity: $T_z = bZ$ relates home production time to Z-good production by the coefficient b , and $C = aZ$ relates purchased goods to the Z-good. The time constraint is $T = H + T_z + T_s$, where H is hours spent working outside the household. The goods, or budget, constraint is $pC = wH + V$, where V is nonlabor income. Combine the time constraint and the technology characterization to get the full-income constraint: $(w_1 + w_2 T_s)(T - T_s - T_z) + V = apZ$. Now, the consumer maximizes the utility function subject to this full-income constraint. From the first-order conditions, the ratio of marginal utilities of Z-goods and sleep is $U_z/U_{T_s} = (ap + bw)/[w_1 + w_2(T_s - H)]$. The right-hand side of this expression gives the relative prices of Z-good consumption and sleep. The shadow price of the Z-good is familiar from Chapter 3: it is the cost of the purchased goods and the foregone wage cost of the time spent in producing a unit of Z. The price of a unit of sleep is the wage rate minus any addition to labor income that occurs as a result of extra sleep on productivity.

The behaviorally interesting aspects of this examination of sleep come from the effects of the various cost components of sleep on how much of it is "chosen." Begin with the effect of nonlabor income on the amount of sleep chosen: $\Delta T_s/\Delta V = \{U_{ZZ}[w_1 + w_2(T_s - H)] - U_{ZT_s}(ap + bw) + bU_{ZT_s}w_2\}D^{-1}$, in which $D < 0$.¹³ The first term, an income effect on demand for sleeping time, is positive ($U_{ZZ} < 0$, as is D). If sleep enhances the utility of consumption ($U_{ZT_s} > 0$), the second term also is positive. The third term is nonzero only if sleep raises productivity outside

the household, and in such cases it works in the opposite direction to the first two terms. To explain this last term, market goods purchased with the extra labor income derived from enhanced productivity can raise utility only if household working time, T_z , is also increased; the increase in T_z , combined with the tendency for sleeping time to increase in response to the additional income, implies an unambiguous reduction in out-of-household working time H . Also, from the ratio of marginal utilities shown in the previous paragraph, if sleep is productive, a fall in H raises the cost of sleep. Thus, this last term in $\Delta T_s/\Delta V$ is a second-order substitution effect working against the pure income effect of the first two terms.

Next consider the effect of the base component of the wage on sleeping time: $\Delta T_s/\Delta w_1 = (U_z - bU_{T_s})(ap + bw)D^{-1} + H\Delta T_s/\Delta V$. This is the usual expression from demand theory, dividing the response in quantity demanded of a good or service into substitution and income effects. The first term differs from the usual substitution effect for a good because it includes the term $-bU_{T_s}$, which appears because the change in w_1 also changes the price of Z-goods. But the full quantity $U_z - bU_{T_s}$ will always be positive, so with $D < 0$, the substitution effect of a wage change on sleep is always negative: the wage goes up, sleep goes down. However, the substitution effect of an increase in w_1 on waking household time, T_z is positive because of the complementarity between T_z and C in the production of Z-goods: as noted above in the discussion of the income effect on sleep, the goods purchased with the extra income derived from the productivity effect of sleep can be enjoyed only if home production time required to process them into Z-goods is increased at the rate aZ per unit of C . Nevertheless, the substitution effect of an increase in w_1 on total home time, T_z and T_s , is negative, just as in the standard labor-supply model of subsection 10.3.1. If the income effect on sleep is small, a higher nonhousehold wage will reduce time spent sleeping.

Suppose the productivity component of sleep increases: $\Delta T_s/\Delta w_2 = T_s\Delta T_s/\Delta w_1 - (ap + bw)U_zHD^{-1}$. The first term will have the same sign

as $\Delta T_s/\Delta w_1$, and the second term is unambiguously positive, so an increase in w_2 will have either a larger positive effect on sleeping time (if $\Delta T_s/\Delta w_1 > 0$, which is not especially likely since it would require the income effect outweighing the substitution effect) than an equal increase in w_1 , or a smaller negative effect.

Returning to the expression for U_Z/U_{T_s} , the demand for sleep could be affected by changes in any exogenous factors that affect tastes for either sleep or Z -goods. Any exogenous factor that shifts the demand for commodities will affect the demand for sleep and the corresponding, residual supply of hours to work outside the household.

An empirical estimate of the elasticity of sleeping time with respect to the full wage rate, w , based on survey data from the United States, is -0.042 , at a high level of statistical significance. For men, an increase in the wage has little effect on H but does move time from sleep to waking household time, T_Z . The estimated income effect on T_s is small, possibly because of the secondary substitution component of the income effect (the third term in the expression for $\Delta T_s/\Delta V$). The principal effect of extra nonlabor income may be to shift the typical consumer to production of relatively more goods-intensive Z -goods through the purchase of more C instead of reducing the total time spent producing those commodities.

Variations in sleeping time over the agricultural cycle in ancient farming societies could have been quite large, and may very well have varied by sex and age. This in turn may have influenced the intrahousehold allocations of food, which, as will be reported in section 10.7, has been observed to affect nutrition and health of children, particularly preadolescent girls, during agricultural peak labor demand periods. Whether sleeping time of children, and consequently health at formative ages, would have been similarly affected is an open question; contemporary evidence sheds no light on that question. I make no claim that the economics of ancient sleep behavior is a critical topic in the analysis of ancient Mediterranean societies, but I hope that the demonstration that the household

production model can shed theoretical insights into a large block of time allocation not commonly considered a subject of rational choice, and that empirical evidence supports the existence of these effects in contemporary societies, may give readers some confidence that they understand how these applications work when we use the same basic model in subsequent sections of this chapter to study decisions about marriage, child bearing and mortality, treatment of children and other topics not commonly thought of as economic decisions. The household production model is able to take subjects ordinarily not thought of as being within the ordinary scope of “economic” decisions and show how the actions involved “cost” time, how time devoted to one activity comes out of some other activity or activities, and how, by the time all the reallocations are finished, there is a genuine “economics” of these subjects.

10.4 Labor Demand

The other major determinant of how much labor is used in an economy, and the price it commands, is of course, the demand side of the matter: how users of labor (employers of others and self-employers) decide how much labor to use according to its productivity and cost. The demand for labor – for any factor of production, for that matter – is governed by both technological characteristics of production and conditions of demand for products. Nevertheless, in studying the relationships between product demand and factor demand, we need not trace the influences all the way back to individuals’ utility functions. Consequently, labor demand is fundamentally simpler than labor supply, although it is subject to the alternative viewpoints offered by looking at production from the direct production perspective or from the dual perspective of costs, or by the responsiveness of employment of various factors to price changes as contrasted with the response of one factor’s price to variations in the availability of the same or another factor.

Production and cost theory do reach their own heights of elegant and powerful abstraction

(translated: difficult, esoteric, gives the intuition a severe workout), but we need not approach those territories to produce some useful insights about influences on the demand for labor. The amount producers are willing to pay for labor depends on the productivity of labor, which in turn depends on other things than the labor alone. The first subsection deals with the firm's demand for labor and how that is affected by choices the firm can make.

The demand for labor is essentially a derived demand – derived from the demand for goods and services produced in part with labor. The second subsection returns to the laws of derived demand introduced in Chapter 2 and uses them to yield some theoretical predictions about empirical magnitudes of relationships that may have relevance to the strengths of ancient relationships. The concepts of these two subsections offer methods for thinking about how employment, either of all labor together or of specific types of labor that aren't perfect substitutes for one another (for example, young and old workers, males and females – for some types of tasks – skilled and unskilled workers, and so on) could be expected to respond to various types of events: changes in demands for products, changes in technologies, changes in numbers of workers of particular types. Should we expect employment to rise or fall, wages to rise or fall? For most people, wages (or payments to labor by any other term) comprise the major portion of income, so these are important issues in the lives of the people under scrutiny.

10.4.1 *The productive enterprise's demand for labor*

It is useful to distinguish between short-run and long-run demands for labor. The short-run is simply the longest time period during which at least one other factor of production is in fixed supply to a producer. In the long run, all factors can be varied by the producer. Let's work initially with the textbook two-factor production function, $Q = f(K, L)$, $f_K > 0$, $f_L < 0$, and $f_{KL} > 0$.¹⁴ Since in the short run, one of the inputs is fixed and we're interested in studying the demand for

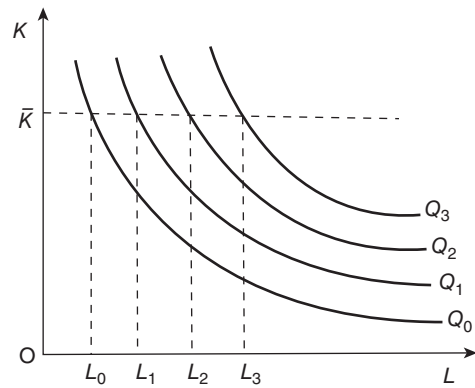


Figure 10.26 Short-run demand for labor by a firm.

labor, that means that we need to specify that the firm using this production function is working temporarily with a fixed quantity of capital – call it $\bar{K}$. Figure 10.26 shows an isoquant diagram with capital and labor on the axes. Recall that along each isoquant, the producer can produce a constant quantity of output Q by varying the combination of capital and labor. But our producer is stuck with $\bar{K}$ capital right now, and his choices boil down to how much labor he's going to want to use. By choosing a quantity of labor services, he is choosing a scale of operation – that is, a quantity of output to produce. How does he choose which quantity of labor to pick?

The optimal quantity of labor in this situation depends on the relative price of labor and capital, which can be described by the negative of the slope of straight lines going from northwest to southeast in Figure 10.26 (which we haven't drawn). Visualize the slopes of such lines at each intersection of $\bar{K}$ with an isoquant; you'll see that the relative price of labor must get cheaper at larger scales of production. Why so? Let's recall one of our earliest diagrams – the one of total product, average product, and marginal product of one factor of production, with the quantity of the other (or all others) held constant; we reproduce this in Figure 10.27. Now, go back to the production function and let's set up the firm's profit maximization problem: $\text{Max}_L pf(\bar{K}, L) - r\bar{K} - wL$. As capital is fixed, the

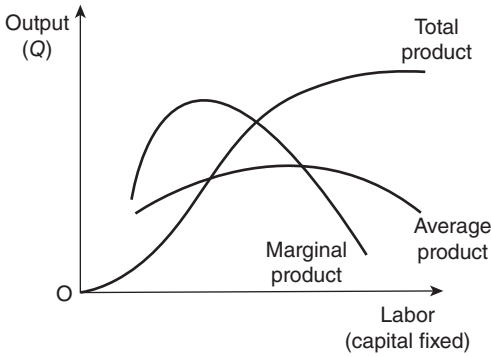


Figure 10.27 Total, average, and marginal product of labor with fixed capital stock.

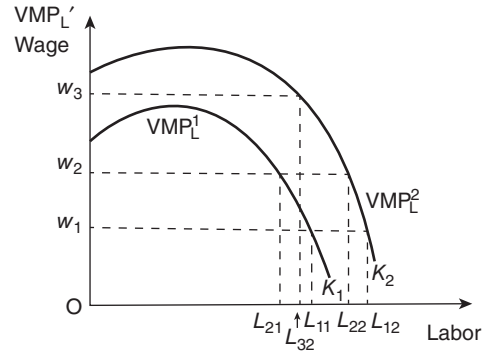


Figure 10.28 Value of marginal product of labor with different capital stocks.

producer's only option for maximizing profit is to adjust the labor input; hence the subscript L after the "Max." The first-order condition on the labor choice is $pf_L = w$: the value of marginal product of labor (sometimes called the marginal revenue product of labor) should equal the wage rate, which amounts to saying that the relative price line in Figure 10.26 should be tangent to an isoquant.¹⁵ Proceeding, we need to recognize that the marginal product of labor, f_L , is a function of both the quantity of capital and the quantity of labor, so we could write it as $f_L(\bar{K}, L)$ or just $f_L(L)$ for short. We can invert that function to get the quantity of labor associated with any particular marginal product of labor, given the quantity of capital in use: $f_L^{-1} = p/w = L^d$, which says that the demand for labor is inversely related to the wage rate and positively related to the product price. Thus the demand curve for labor is the downward-sloping portion of the marginal product of labor curve multiplied by the product price. The basic form of the labor demand curve is $L^d = f(w)$, with the property $\Delta L^d / \Delta w < 0$: "demand for labor is a (negative) function of the wage." The analogy to the demand curve for any consumer good is exact.

For equilibrium in the labor market, we need the labor supply, from subsection 10.3.1, to be brought into equality with the demand for labor, by adjustment of the wage: $L^d = f(w) = L^s = g(w) = L$, in which L^s is the labor supply curve, with $\Delta L^s / \Delta w > 0$.

Figure 10.28 shows these value of marginal product (VMP) curves for labor, with different quantities of fixed capital in use. When capital in the amount K_1 is used, labor's VMP curve is VMP_L^1 . At wage w_1 , a producer would want to hire labor in the quantity L_{11} . If the wage rose to w_2 and the producer still used K_1 capital, he would want to cut back on employment to L_{21} . If the wage continued to rise to w_3 , if the producer could not alter his capital stock, he would want to shut down operations because it would cost more to hire labor than the labor could produce. If the employer had K_2 quantity of capital (greater than quantity K_1), he would want to hire L_{12} labor at wage w_1 , because the additional capital has increased the productivity of labor (remember the cross-effect $f_{KL} > 0$). Similarly at w_2 : with K_2 capital, this producer would want to hire more labor (L_{22}) than was profitable with capital stock K_1 . If the wage rose as high as w_3 , he would find it still profitable to operate, although with labor $L_{32} < L_{22}$. These curves showing the quantities of labor that a producer would want to hire at various wages are labor demand curves. They are functions of the wage, the product price, the prices of other factors, and the quantities of other factors in fixed supply: $L^d = L(w; p, r, \bar{K})$, with the response properties $L_w < 0$, $L_p > 0$, $L_r > 0$, and $L_{\bar{K}} > 0$. When the wage rises, the quantity of labor demanded falls – it slides down a given VMP curve. When the product price rises, the VMP_L curve shifts out and the quantity of labor

demand rises, assuming everything else stays the same. When the rental cost of capital rises, the VMP_L curve shifts out, and if the quantity of fixed capital increases, the VMP_L curve also shifts out. Be careful to distinguish the difference between the quantity of labor demanded changing along a fixed demand curve for labor (VMP_L curve) and the demand curve for labor itself shifting, with the quantity of labor demanded along the new demand curve most likely changing too. The quantity of labor demanded can change without any change in the demand curve for labor; a change in the wage rate is sufficient to induce such a change.

Now let's turn to the long-run demand for labor, when all other factors can be varied as well as labor. Figure 10.29 will help us tell the story of factor substitution in the face of factor price changes. This is the familiar isoquant diagram, beginning with output level Q_1 and the factor price ratio represented by the steep line with vertical intercept at c_1/r_1 and horizontal intercept at c_1/w_1 , where c_1 is the total production cost of output Q_1 at the factor price combination (w_1, r_1) . Labor in the amount L_1 will be used to produce Q_1 (we ignore the quantities of capital used). Now, let the wage fall to w_2 , with the rental price of capital remaining at r_1 . Maintaining the original output will let total production cost fall to c_2 , which has the effect of sliding the factor price line along the isoquant, just as we have studied substitution and income effects in consumption problems and substitution and output effects in production

problems. The increase in employment from L_1 to L_2 is a pure substitution effect, but this reduced cost of producing the original output means that the firm whose isoquant we are studying now can produce the product for less than its market price. Consequently it has the incentive to expand production to earn the profits $p - c_2$ on more units than Q_1 . Sadly, of course, it drives up its costs in this expansion, all the way back to $c_3 = p$ at output level Q_2 . In undertaking this cost-increasing expansion it reaches a tangency with isoquant Q_2 and the factor price line with horizontal intercept at c_3/w_2 , where it employs L_3 labor. The expansion from L_2 to L_3 is the scale effect. We now show some precise measures of these changes.

The increase in employment from L_1 to L_2 involves the substitution of labor for capital, induced by a reduction in the price of labor relative to the price of capital (remember that the rental price on capital, r_1 , didn't change). The percentage change in the factor-utilization ratio, K/L , following each percentage point change in the inverse factor price ratio, w/r in this case, with output constant, is the elasticity of substitution (commonly designated σ , and always measured positively when there are two inputs).¹⁶ We can offer several measures of this. The intuitive one we just gave is $\sigma = \% \Delta(K/L) / \% \Delta(w/r) \geq 0$. You can see that the capital / labor ratio in Figure 10.29 falls from the factor combination at L_1 to that at L_2 . Another measure of the elasticity of substitution, based on characteristics of the production function is $\sigma = f_{KL} f_L / Q f_{KL} > 0$. These are equivalent expressions.

With more than two inputs, pairs of inputs can be either substitutes or complements. If the constant-output σ_{ij} is positive, inputs i and j are called substitutes in production; if negative, they are called complements in production. If a pair of inputs are complements in production, a decrease in the price of one of them should lead to an increase in the use of the other – just the opposite of what happens between pairs of substitutes in production. This particular elasticity of substitution is called a partial elasticity of substitution because it holds output and other factor prices constant, considering the interaction between the designated pair of inputs only. The

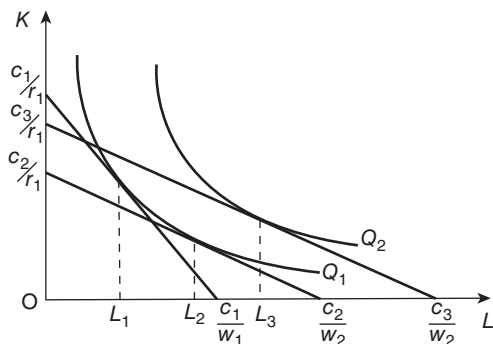


Figure 10.29 Long-run demand for labor by a firm.

production-function based formulation for σ_{ij} is more intricate than that for the two-factor case, but its interpretation as the percentage change in a factor utilization ratio per 1% change in the inverse factor price ratio, is the same.

Now we move from the elasticity of substitution between inputs to elasticities of demand for labor – the “own-price” elasticity and cross-price elasticities. The own-price elasticity of demand for labor for the two-input case, with output constant is $\eta_{LL} = -(1 - s_L)\sigma < 0$, where $s_L = wL/pQ < 1$ is the share of labor in production cost (with constant returns to scale in production, the capital share, $s_K = (1 - s_L)$). The cross-price elasticity, that is, the percentage change in labor employed per 1% change in the rental price of capital, is $\eta_{LK} = (1 - s_L)\sigma > 0$. If we add the scale effects of moving from isoquant Q_1 to Q_2 in Figure 10.29, the own- and cross-price elasticities of demand for labor are $\eta'_{LL} = -(1 - s_L)\sigma + s_L\eta_Q$, where $\eta_Q < 0$ is the elasticity of demand for the product, and $\eta'_{LK} = (1 - s_L)(\sigma + \eta_Q)$. The scale effect can be said to be the short-run component of the demand for labor, since no production rearrangements involving factor substitution have to be incurred by the producer – just increases or decreases in output using the same technological choices. In the long run, producers will make the changes in technology choices that involve substituting between factors.¹⁷ This brings in the substitution effect as well, so η'_{LL} and η'_{LK} are long-run own- and cross-elasticities, containing both scale and substitution effects.

Pairs of inputs with negative total cross-elasticities are called “gross substitutes”; those with positive cross-elasticities are said to be “gross complements.” Whether a pair of inputs are gross substitutes or gross complements depends on the relative magnitudes of the substitution and scale effects. If two inputs are substitutes in production, they can be either gross substitutes or gross complements, depending on whether the substitution effect is larger or smaller than the scale effect. Pairs of inputs that are complements in production must be gross complements: the scale effect necessarily will move in the same direction as the movement along the isoquant. “Gross complement” and “gross substitute” refer to the sign of the scale-inclusive cross-elasticity of

demand between two inputs, η'_{ij} . “Complement in production” and “substitute in production” refer to the sign of the partial elasticity of substitution, σ_{ij} . The principal intuition between the concept of pairs of factors being substitutes is that when the price of one of these two factors rises, the quantity employed of the other rises; whether we allow for changes in scale of output when we make this observation tells us whether we are thinking about “gross substitutes” (scale changes) or “substitutes in production” (scale is unchanged). Correspondingly, when a pair of factors are complements, when the price of one of them rises, the quantity employed of the other goes down, and we have the same qualifications about whether scale changes or not when we make the observation about the employment change.

So what do these concepts of substitutes and complements in production and gross substitutes and complements mean in practical terms? Suppose two types of labor – say, skilled and unskilled workers – are substitutes in production. If a technological change increases the demand for skilled workers, a sufficiently large scale effect could make skilled and unskilled workers gross complements, leaving the unskilled workers unhurt by the technological change. Alternatively, if unskilled workers and capital equipment are substitutes in production and some technological change reduces the price of the capital, those workers will be hurt (their wage will fall) if the scale effect is not large enough to outweigh the substitution effect.

These effects on the demand for labor rely strictly on changes in the relevant factor prices. Although changes in product prices will affect the demand for various categories of labor, labor demand elasticities themselves hold product prices constant. To examine the effects of product demand and supply characteristics on labor demand, we turn now to the derived demand concept.

10.4.2 Derived demand

When thinking of the derived demand for labor or for any other factor of production, we need to think in terms of the demand coming from a specific employment of that factor, for example,

in shipping, or in agriculture, or in ground transportation such as teamstering. While it does make sense to think in terms of an aggregate demand for labor throughout an economy, the derived demand concept links the demand for labor in a particular employment to a specific line of production. The expression we developed above for the long-run labor demand elasticity, $\eta'_{LL} = -(1 - s_L)\sigma + s_L\eta_Q$, is a simplified version of the elasticity of derived demand for labor – under the special assumption that the supply of the other factor used in production (for simplicity, suppose there are just two factors) is in perfectly elastic supply (that is, the supply elasticity of the other factor – call it capital – is infinite). What this particular expression tells us is that the more elastic is the demand for the product – the more the demand for the product changes as its price changes – the more elastic will be the demand for labor used in producing it, given any degree of substitutability between capital and labor and the share of labor in production costs. (Note that the derived demand elasticity is an own-wage elasticity – that is, the elasticity of demand for labor with respect to its own wage rate – not an elasticity of employment with respect to the price of the final product.) It also tells us that the larger the share of labor in production costs is, the more elastic is the demand for labor: with a larger labor cost share, a given percentage increase in the wage would force production costs up by a larger amount than would occur with a smaller cost share, for any degree of substitutability between labor and capital. Think about this latter point slightly differently: if labor's cost share were very small, a fairly large increase in the wage would not be particularly noticeable by consumers in the form of a product price increase, so there would be a smaller reduction in employment, if there were any at all – as long as the product demand is more elastic than are substitution possibilities in production. This is called “the importance of being unimportant,” because it implies that factors with small cost shares can bring relatively extortionate wage increases from employers, simply because in the overall scheme of things, it doesn't cost the employer much that is noticeable on the final product end.

In the more general case, the supply curve of the other factor will not be perfectly elastic. The general formulation of the derived demand elasticity is considerably more intricate than the special case we first developed, although its derivation is essentially quite simple – a series of “differencings” of the firm's revenue constraint, $pQ = wL + rK$. So far, we have measured the elasticity of substitution so that it is always positive (if not zero), and of course, the own-wage elasticity of demand for labor is negative – the quantity of labor demanded falls when the wage rises. However, simply by the conventions that have arisen in the theoretical development of the elasticity of derived demand, all the elements of that elasticity are measured positively – the elasticity of substitution, the elasticity of product demand, even the derived demand elasticity itself.¹⁸ Let η , measured positively, be the negative of the product demand elasticity; $\varepsilon > 0$ be the supply elasticity of the other factor (call it capital); $\sigma > 0$ be the elasticity of substitution in production between labor and capital, and s_L be the share of labor in production cost. These are the four basic parameters underlying the derived demand for labor (or any other factor), whose own-wage elasticity we will designate as λ , which is also measured positively. Then $\lambda = [\sigma(\eta + \varepsilon) + s_L\varepsilon(\eta - \sigma)]/[\eta + \varepsilon - s_L(\eta + \sigma)]$. If the supply elasticity of the other input is infinite, the expression simplifies to $\lambda = (1 - s_L)\sigma + s_L\eta$, which is simply the expression with which we opened this subsection, with the signs reversed. If the other factor is in perfectly inelastic supply ($\varepsilon = 0$), it simplifies to $\lambda = \sigma\eta/[s_L(\sigma - \eta) + 1]$, or in an interesting inverse form, $1/\lambda = s_L/\eta + (1 - s_L)/\sigma$, which breaks cleanly into scale and substitution-in-production effects.

The derived demand elasticity has three unambiguous properties, and one other that is contingent. These are commonly called Marshall's four laws of derived demand. First, $\Delta\lambda/\Delta\sigma > 0$: the derived demand for labor is more elastic the more elastic is the substitutability in production between labor and the other input. With zero substitutability (the case of the fixed-coefficients production function), there is no reduction in employment via the substitution effect; only the

scale effect is available to accomplish the required adjustment, so reductions in employment caused by a rising wage would not be mitigated at all through substitutions. The limitations on substitutability need not be purely technological, however. Legal or other institutional restrictions can restrict, particularly, substitutions *away* from labor when its wage rises.

Second, $\Delta\lambda/\Delta\varepsilon > 0$: the derived demand for labor is more elastic the more elastic is the supply of the other factor. When a wage increases and employers try to substitute into other factors and away from labor, they drive up the demands for those other factors. How quickly and by how much the shift into other factors will raise their prices depends on their supply elasticities. If there is a fixed supply of the factor into which the employer tries to substitute for labor, the price of that factor will be bid up quickly as employers try to increase its use (which, of course, they can't do – they can only drive up its price). Such substitutions are less likely to bid up the prices of other factors in the long run than in the short run; that is, although the other factor's price may be bid up in the short run, in the long run, when its supply can be augmented more fully, its price will come back down, and the employers trying to substitute into it to avoid higher cost labor will be able to do more substitution. This effect is another reason why long-run labor demand elasticities are larger than short-run ones.

Third, $\Delta\lambda/\Delta\eta > 0$: the derived demand for labor is more elastic the more elastic is the demand for the final product. Because of this relationship, the elasticity of demand for labor at the level of the individual employer is greater than that at the level of the industry and the market.¹⁹ This law also implies that wage elasticities will be higher in the long run than the short run, once again because there are more product substitution possibilities in the long run.

Finally, $\Delta\lambda/\Delta s_L > 0$ only if $\eta > \sigma$, product demand is more elastic than is factor substitutability; it is negative if $\eta < \sigma$ and zero if $\eta = \sigma$: the derived demand for labor is more elastic the larger is labor's share in costs only if consumers of the final product can substitute away from

this product if its price rises more easily than producers of this product can substitute away from labor if its wage rises. This is the qualified "importance of being unimportant." Intuitively, if labor's cost share in production is large, when the price of labor rises, the production cost increase is relatively large, and the producer will have to reduce employment to contain the cost increase; the employment reduction is larger the larger the cost share, but is mitigated to the extent that the producer can substitute away from labor. When labor's cost share is already small, there is little scope for substituting away from it when its wage rises, hence the opportunity for such a small-share factor to raise its price by some type of collective action. This is one reason that unionization is more effective in raising a unionized wage when only a few are covered by the union wage: their cost share is small, leaving little room to substitute away from them.

Derived cross-elasticities – for example, the elasticity of demand for capital with respect to the wage rate of labor, or the elasticity of demand for skilled labor with respect to the price of unskilled labor – involve the price of one factor changing and the employment of another factor responding. Consequently Marshall's four laws of derived demand can't be applied directly, but each of the four parameters corresponding to one of those laws does affect the magnitudes of the scale and substitution responses and correspondingly offers insights into the overall employment response. The general form of the derived cross-elasticity, for example the elasticity of demand for capital with respect to a change in the wage of labor, is $\mu = s_L(\eta - \sigma)\varepsilon/[s_L(\sigma - \eta) + 1 - \varepsilon]$, which simplifies to $\mu = s_L(\eta - \sigma)$ when the supply of "the other factor" (in fact, the one whose price has changed) is perfectly elastic. Unlike the derived own-wage elasticity, λ , the derived cross-elasticity, μ , can be positive, negative, or zero. Note that the cost share, s_L , refers to the cost share of the factor whose price is changing, not the one whose level of utilization is changing. The supply elasticity of "the other factor" now refers to the factor whose employment is changing. Consequently, when $\varepsilon = 0$, it makes no sense to calculate the demand elasticity of the factor other than labor,

because its quantity employed will not respond. For this reason, when $\varepsilon = 0$, we calculate the inverse of the previous cross-elasticity, which gives the response of the price of capital to a change in the quantity of labor available, which is $1/\mu' = s_L(1/\sigma - 1/\eta)$.²⁰ The price and quantity cross-elasticities are related to each other by the relationship $\mu\mu' = \lambda\varepsilon$. We address how each of the four parameters affects the magnitudes of the scale and substitution effects.

When the factor whose price is changing (we've let that be labor above) accounts for a larger production cost share, there is more scope for large product price changes and a correspondingly large scale effect. Consequently, there is a greater likelihood that the two factors will be gross complements when the share of the price-changing factor is large. The demand elasticity for the product also has an important influence on the scale effect: the larger is η , with other things equal, the larger the scale effect will be.

The larger is the elasticity of substitution, σ , the easier it is to substitute between the factor whose price has changed and the one whose use we are tracking with the derived cross-elasticity. If the two factors are complements in production, the effect on the use of the other factor, when the price of labor changes, reinforces the scale effect and always increases that factor's employment – making the pair of actors gross complements as well as complements in production. If the two factors are substitutes in production, a wage change will produce an opposite movement in the employment of the other factor, and the question becomes whether the substitution effect is large or small relative to the scale effect. A more elastic supply of the other factor leads to a larger substitution effect in cross-elasticities. While a larger cost share of the factor whose price changes produces a larger scale effect, it makes for a smaller substitution effect, simply because there is little scope for substitution in the factor with the smaller share.

Now, let's go back to our usual sign conventions about own- and cross-price elasticities – measuring own-elasticities negatively and cross-elasticities either positively or negatively – and think about the full set of elasticities for all

the factors in a multi-factor production process. Suppose the production function for the firm or industry we have in mind is $Q = f(L_1, \dots, L_m, X_1, \dots, X_n)$, where the L_i are various types of labor (say, various skills, plus unskilled workers) and the X_i are other inputs such as different types of equipment, animals, and possibly land. We have $m + n$ factors and the same number of factor prices. For each factor price change, we have $m + n$ responses, or elasticities, for example, for the change in the wage rate of labor of type 1, say carpenters, the full set of responses to a change in the carpenters' wage. For a constant-returns-to-scale production process, the sum of the elasticities has to equal zero: $\eta_{L_1 L_1} + \eta_{L_2 L_1} + \dots + \eta_{L_m L_1} + \eta_{X_1 L_1} + \dots + \eta_{X_n L_1} = 0$. Since we know that the own-wage elasticity is negative and that the sum of all the elasticities is zero, at least one of the cross elasticities is positive: at least one pair of inputs are substitutes. This in itself is not particularly startling, but it leaves room for several $\eta_{ij} < 0$, which implies complements, in the multifactor case. Additionally, the signs of the cross-effects are symmetric – that is, if $\eta_{ij} > 0$, then $\eta_{ji} > 0$, but generally $\eta_{ij} \neq \eta_{ji}$ because the cost shares of factors i and j will differ.

We have introduced the distinction between the elasticity of an employment level as factor prices change and the elasticity of a factor price as factor availability changes. Substitutability and complementarity can exist between pairs of factors in terms of either type of elasticity. These relationships with the former type of elasticity – the more commonly used one, which measures the responsiveness of employment to factor price – are called “ p -substitutability” and “ p -complementarity.” They are the type of substitutability and complementarity we have already discussed relatively extensively, both as “in production” and “gross.” Using this concept of substitutability and complementarity, factor prices are exogenous and employment levels are endogenous. If skilled and unskilled labor are p -substitutes, an increase in the skilled wage increases the proportion of unskilled workers in production. With the other type of elasticity, we have what is called “ q -substitutability”

and “ q -complementarity.” Factor quantities are exogenous while factor prices respond endogenously. In general, we measure these q -elasticities as $\varepsilon_{ij} = \% \Delta w_i / \% \Delta L_j$, where subscripts i and j may refer to different types of labor, but need not (that is, $i = j$ is an option, in which case we are talking about an own-elasticity rather than a cross-elasticity). (Also, don’t confuse ε_{ij} in this notational use with the other-factor supply elasticity ε used in the derived demand elasticity formulas.) If $\varepsilon_{ij} > 0$, the elasticity of the wage of factor (labor type) i with respect to the availability of factor (labor type) j , inputs i and j are q -complements: an increase in the availability of labor type j raises the wage (marginal productivity). Recalling that an increase in the marginal productivity of a factor is equivalent to decreasing its proportion relative to other factors, a pair of factors could be p -substitutes and q -complements if, say, an increase in the employment of skilled workers raised the relative scarcity of unskilled workers.

Factor demand elasticities and elasticities of substitution in production are highly industry- and technology-specific, but some useful benchmarks for ancient possibilities may be provided by summarizing some contemporary evidence from Hamermesh (1993 Chapter 3) on magnitudes of these economic parameters. First, across a wide variety of industries and countries, the short-run elasticity of demand for labor ranges from around -0.15 to -0.75 , with a central tendency around -0.30 . The scale effect ranges probably from around -0.55 to possibly as high as -0.70 . Substitution elasticities in production between labor and capital (σ) average around 0.45 . Most estimates of the long-run elasticities of demand for labor are less elastic than -1.0 . From studies that disaggregate more among inputs, labor and materials are p -substitutes, with a small cross-elasticity; similarly with labor and energy. Capital and energy (the latter provided to a considerable extent by humans in antiquity) are probably p -complements in contemporary economies, and they are jointly p -substitutable for labor. Labor, without reference to different characteristics, is generally a substitute for each of the other major input aggregates for which

production and cost functions are estimated (capital, energy, and materials). However, skilled labor and capital are probably p -complements, and additional training of labor reduces the degree of p -substitutability between labor and capital and generally reduces the own-wage elasticity of demand for labor (η_{LL}). The demand for skilled labor falls relative to that for unskilled labor as capital equipment ages, as more people have the opportunity to learn how to use it. Unskilled labor is likely to be a p -substitute for capital (implying symmetrically that capital is a substitute for unskilled labor). Skilled and unskilled labor are substitutes in production. There is strong evidence for capital-skill p -complementarity. In work on American cotton farms in 1860, the demand for free labor was less elastic than that for slave labor. Finally, complementarity or substitutability between immigrants and native workers, and between recent and earlier immigrants is small, implying that immigrants have had relatively minor effects on wages of native workers.

The substitutability between labor and materials – you can use “more” labor (for example, skilled labor) and fewer materials or less labor (less skilled labor) and more materials – has interesting implications for ancient evidence. Evidence from Melos, in the Cyclades, of periods of greater and lesser wastage of obsidian has been interpreted as evidence of (degrees of) alternative periods of monopoly and competition among artisans (Torrence 1986, Chapter 3). It may, in fact, be better evidence of the average skill level – or the level of demand for obsidian.

The examples of the differential relationship between capital and skilled and unskilled labor also offer interesting benchmarks for ancient relationships. During periods in which technology may have changed little, the productivity advantages of many skills may have been low, considerably dampening the incentives to obtain training. With continuing low rates of technical change over a sufficiently long period, institutions for training would have been quite weak, with few economies of scale available at sizes larger than the family or the individual tutor. The existence and curricula of large, organized schools at

various times and places in the ancient Mediterranean region warrants investigation. Harris (1989) has considerably advanced the study of schooling in ancient Greece and Rome, and Caruso (2013) has analyzed possible archaeological remains of the Academy ("Plato's Academy") in Athens, yielding, *inter alia*, suggestions of size (scale) of a classroom building.

The degree of substitutability between labor and energy may have been much stronger in antiquity than at present, particularly in light of low energy conversion efficiencies of much of the ancient capital stock, including animals used for traction. The low degree of substitutability between contemporary immigrants and natives may be attributable to both language barriers and training differences. In antiquity, immigrants were frequently slaves, and their degree of substitutability for free labor has been implied to have been considerable, particularly in the Roman Empire. Nonetheless, free immigrants – including but not restricted to refugees and forced migrants who retained some status of freedom under the Neo-Assyrian Empire – were not particularly rare in many Mediterranean basin regions in antiquity. Consideration of their effects on the productivity and wages of natives, and how that affected the outlooks on migrants might be a topic that would repay close attention.

10.5 Labor Contracts

Anytime an employer hires the services of some person other than himself, he and the employee create a contract between themselves. It may be a passing oral agreement, it may have all the appearance of what we today sometimes call custom, it may be written in considerable detail. The dimensions of work are manifold, and there is the ever-present distinction between hours and effort which offers the worker returns from shirking; both call for monitoring, which may be difficult or impossible, depending on the tasks involved. People have worked out different kinds of arrangements to solve or acceptably mitigate these problems. These arrangements involve different bases for pay, different timings during the anticipated period of employment,

and a number of implicit contracts that must be in the best interest of both parties to honor. The payment schemes described in this subsection are used in contemporary economies. Some aspects of them, such as retirements and pensions, may seem particularly time-bound and irrelevant to the ancient Mediterranean societies. Possibly. Possibly not. The concepts presented here can take a variety of forms, and it is the structures, not the names, that are important. Even if some contemporary payment arrangements turn out to have been absent in those early societies, the circumstances of their absence may prove informative.

10.5.1 Information problems and incentives

Employment contracts are principal-agent relationships. The effort and attention that an employee puts into a job frequently are only partially observable, and then at a time too late to make modifications for the current production cycle. The supply of labor is governed by the utility of leisure and the utility of goods that can be purchased with the results of working. Consequently, most workers are going to find, now and then, tasks they would rather not do if they can get away without doing them – tasks that their employers are paying them to do but may not be able to monitor sufficiently for enforcement. When the effort being put into a task need not be discernible for some difficult-to-predict length of time, the employer does not have a simple monitoring device – time on the job need not correlate well with effort. Depending on the characteristics of the particular job, some bases of pay (for example, time, piece rate) can generate moral hazard problems. For instance, a guaranteed wage can lead to shirking.

Most employment contracts are incomplete, in the sense that they can't possibly spell out every specific task the employee may be required to do. Consequently many "clauses" remain implicit – legally unenforceable but "incentive compatible" if they are to be effective. That is, honoring one's part of the implicit contract is in both parties' best interest.

Both employers and employees commonly have information about themselves that the other finds difficult to discern, at least not until well into an employment relationship. Potential employees have an incentive to say that they are hardworking, and potential employers have a corresponding incentive to understate the difficulty of the job or overstate its benefits. "Join the Navy and see the world!" could in actuality mean "spend four years in dry dock at Norfolk." "Join the Army and you could become a radio repairman, a skilled vehicle mechanic, an Army bandsman, a medical corpsman, you name it! But chances are a million-to-one you'll wind up in the infantry, buddy!" Employers may be able to elicit signals from potential workers that substitute actual behavior for statements about themselves. As we discussed in Chapter 7, the offer of alternative contracts may give workers the opportunity to reveal some characteristic that the employer believes important, such as initiative, work ethic, or skill. On the other side, employers develop reputations for truthfulness, which can affect the wage they must offer to secure any particular skill level of worker.

Alternative bases and timing of payment and the potential for dismissal can be packaged in ways that align the incentives of workers and employers and either resolve moral hazard problems or reduce them to acceptable dimensions. The next two subsections discuss options for paying workers, both at a particular point in time and over a longer period of time. The differences in working conditions that various jobs may offer – ranging from location to discomfort and outright danger – means that different jobs are package deals – they contain many aspects – and the total wage paid may be compensation for aspects of working conditions that vary across jobs. The last subsection discusses the payment differentials that compensate for such comforts and discomforts.

10.5.2 *The basis of pay*

The alternative bases of payment for labor services are time and output, but a number of hybrids can be constructed. A constraint on the development of alternative payment bases is that the marginal

product of labor should not fall below the wage payment. If it did, the capital and other inputs used by the producer (the employer) would fall below the returns they could earn elsewhere, and they eventually would exit such a producer's firm and seek use where their returns were higher. This restriction, of course, operates on the condition that the other factors being used actually have alternative employments. How might this factor mobility work?²¹

Take a farm for an example. The "employer" in this case is the farm owner (the landlord), and the employees are hired farm workers. If the landlord pays the workers more than their marginal product, part of that payment is going to come out of the payment to land. The landlord will be able to see how much he gets to keep of the value of the output (the return to the land; ignore the returns to his own managerial inputs for the moment). If it's less than what the land could offer him when the labor was paid its marginal product, some other farmer could offer him the land's marginal product to purchase the farm, and it'd be more (in present discounted value terms) than he could get from operating it himself. He's then faced with the dilemma of continuing to operate the farm and earning less than he could by selling it. Setting aside other personal or cultural restrictions on land sale (which may be overrated as to their unexplainability in resource allocation terms anyway), he sells, and the wage payment to labor falls back to its value of marginal product. Take another example: a bronze smith who rents his equipment. He pays his helpers more than their marginal product. He still has to pay prices equal to the marginal product (equals marginal cost) of his variable inputs – firewood, scrap bronze, any alloy material – and rent equal to the marginal product of his equipment – bellows, tongs, coal shovels, crucibles, molds, and so forth. As the operator, he takes the residual earnings as his own wage. When he gets finished paying everybody he owes, what is left for him is less than he could earn elsewhere; if not, he goes in debt to one of his suppliers, and there's no reason they should tolerate accepting less from him than they could get elsewhere, so they stop selling to him, and he's out of business. Either the bronze smith accepts

a lower wage payment for himself or he shuts down his operation and works for somebody else to earn more than he was paying himself as a self-employed bronze smith.

Payment by time is simply an hourly (or some other time unit) wage, regardless of effort or output. If tasks are simple and easy to monitor, this simple time basis of pay can maintain incentives of workers and productivity for the employer. The salary, of flat rate, is a variant of the straight time payment, but it is not as strictly contingent on attendance as the hourly wage. It is used when effort is hard to observe, output difficult to observe over short periods, individuals have considerable control over the pace of work, and the time to complete a task is uncertain and hard to predict.

The basic form of payment by output is the piece rate. A pure output basis for pay puts all risk of fluctuations on employees, even from sources beyond their control. When tasks suitable for piece-rate payment, such as military recruiting, have been paid by both pure bases, the piece-rate scheme has been found to raise productivity by 15–20%. On the other hand, if an employer paying with a piece-rate scheme found that he had been placing more product on the market than could be sold at going prices, he could find it difficult to reduce output by reducing hours of work: piece-rate workers appear to find the flexibility to increase their pace of work to maintain income. The dispersion of income among workers paid piece rate can be considerable, reflecting both skill and demand for income. Piece-rate payment is not appropriate when haste can damage the product or equipment. All output-based payment schemes require that workers' performance can be measurably related to some objective of the employer. A tavern-keeper's assistant may be friendly and helpful to customers and tavernkeeper alike, and may be a highly valued employee, but the job offers no simple measures on which output-based payment could be implemented.

A variant of the piece rate system is share wage payment, by which workers, either individually or in groups, are paid a share of the output, specified in advance. A combination of their skill and

luck will determine how much they actually earn. Share wages have been used in agriculture in a number of countries. Share wages shift risk from an employer to employees, but can offer a monitoring device to mitigate shirking problems when direct monitoring is costly.

Bonus and profit-sharing systems are hybrids of time and output payment. The bonus is added to salary to reflect the employer's assessment of individual productivity. It is typically used with high-income workers, although the closely related booty system – a profit-sharing system – was used in addition to regular wages in ancient armies and navies (Trundle 2004, 99–101). Payments from profit sharing will appear irregularly and depend on the productivity – or luck – of all workers. Little is known yet about the effects of profit sharing on a firm's productivity.

The choice of payment basis can be tailored to characteristics of individual lines of production so as to mitigate moral hazard problems, to reduce ordinary shirking outside moral hazard situations, to retain desired employees ("reduce turnover" in the lexicon of labor economics), encourage less suitable employees to seek other employment, and generally boost productivity. One possible payment scheme that might be used to raise productivity is called "the efficiency wage hypothesis." It's called a hypothesis because that's its current observational status. The basic idea is that if an employer pays her employees more than marginal product, they will work harder, repaying the higher wage payment. The mechanism depends on the variability of effort, among other factors. There are two variants of the hypothesis, absolute and relative. The absolute efficiency wage concept has been directed at low-income agriculture in contemporary developing countries, on the grounds that if people are paid more they can eat better and will consequently improve their productivity through improved health and stamina. The idea sounds good, and some studies have found associations between payment and effort in developing-country agriculture, but in general whether higher pay will find the hypothesized linkages to health and productivity remains an open question. The relative efficiency wage concept is more general, depending as it

does on effort increases from whatever source when workers are paid more. Workers' value of marginal product (VMP) actually rises with the wage. Of course, as the wage continues to be raised, the increases in VMP will get smaller; when the marginal increase in the wage equals the marginal increase in VMP, the employer has found the optimal efficiency wage. The sources of effort and other efficiency increases include improved satisfaction and reluctance to leave a job (or be fired from it for shirking) because they make more than they could for comparable work elsewhere.

10.5.3 *Sequencing of pay*

When workers get paid, as well as the basis and level of their payment, can affect their productivity. Comparable to the constraint on the basis of pay, that wage payments be no greater than the value of labor's marginal product, sequencing plans are subject to the constraint that any period of "overpayment" of labor (paying a wage greater than the value of marginal product) be exactly offset by a period of "underpayment." Otherwise, the employer eventually will go broke, as his labor payments gradually eat away the value of his capital equipment. Mobility of labor will operate to maintain the equality in the other direction, that is, keep the discounted value of the wage deficit no greater than the discounted period of the wage surplus.

Deferring some portion of payment is an effective discipline device, equivalent to having workers post performance bonds. In longer-term employment associations, it is common for workers to be paid less than their marginal product in early years on the job and more in their later years. Properly discounted, the excess of the wage over marginal product in senior years should balance the deficit of the wage below marginal product in early years. A component of this payment scheme is an agreement on when the worker will retire, because the excess of wage over marginal product could present an attractive incentive to remain working in older years at the net expense of the employer. Suitable adjustments to pensions can serve as an effective alternative

to binding retirement date agreements. Younger workers will be attracted to remain on the job to reach the time when they are "overpaid," and the disciplinary, or counter-shirking, effect will become stronger as workers age. Older workers will be disciplined by the prospect of losing their overpayment by being fired for shirking. The firm will be discouraged from trumping up charges against older workers so they can fire them because of the costly impact on their reputation: they would have to pay a higher base wage to attract workers who anticipated such treatment in an implicit contract such as this.

One limitation on the use of such sequencing plans would be a young and highly variable, expected mortality age. Most workers would not expect to live to the date at which they start making up for their early underpayment. However, for segments of ancient populations that might have had more secure, and longer life expectancies – people such as upper-class court personnel – such sequencing schemes might have been able to operate to maintain youthful attentiveness and loyalty.

Most long-term employment relationships provide opportunities for advancement up the "job ladder" of skills. The opportunity to advance is a disciplinary device commonly used in conjunction with the straight-time basis of payment. Such advancements have the characteristics of a tournament: absolute performance is less important than relative performance. An employer may find it difficult to observe productivity per se, but ranking employees may be considerably easier.

10.5.4 *Compensating differentials in wages*

The compensating differential concept accounts for the fact that most jobs confer direct consumption of various amenities and disamenities on the workers who hold them – frequently whether they want them or not. The compensating differential is a difference in the directly paid wages between different jobs. There is no easily observable "base" wage against which to measure the wages in all other jobs; sadly, it is an intricate statistical effort to identify these wage

differentials although the intuition underlying the concept is easy to grasp. Suppose we had two shipyards in which workers built wooden ships – somewhere on the Levantine coast, say, in the thirteenth-century B.C.E. In one of them, the workers are allowed to take home with them oddly sawn, “extra” pieces of wood that “wouldn’t fit in anywhere.”²² In the other yard, the workers get paid a higher wage but don’t get to take any wood out of the yard for personal uses. Can we say that the workers at one yard are better off than those at the other? No. Particularly if there were mobility of workers between yards, we would have to presume that, at least at the margin (that is, the marginal worker), the workers are equally well off, just taking their total income in different compositions of shekels and the king’s wood. The difference between the higher wage at the yard that controls the wood closely and the lower wage at the other yard is the compensating differential.

In the example of the shipyards, the workers in the lower-wage yard can exercise discretion over how much wood they remove – effectively, steal. They can deliberately saw perfectly useful pieces into pieces that aren’t useful so they can take them out of the yard – undoubtedly within limits imposed by what the foreman will tolerate. What the foreman will tolerate actually can be given some more precise dimensions: the limit on the wood will be the tradeoff at the margin between how much more the yard would have to pay its workmen in shekels to get them to stop stealing so much wood. The equilibrium is found when the marginal increase in the wage just equals the marginal savings in stolen wood.

The magnitudes, and even the directions, of these wage differentials are based on the individual’s assessment of how his or her utility is affected by tradeoffs between the purchasing power of a wage and the concomitant consumption of something else that necessarily accompanies a wage of a given level. But as the shipyard’s perspective on the matter brings up, preferences are not the complete explanation of the magnitudes of compensating wage differentials. The other part of the story is how easy it is for the firm in question to control the amenity or disamenity – “easy”

being measured in terms of how much it costs them, which they could turn around and offer directly in wages. Since we know that payment in a fungible good like cash is more efficient than most payments in kind, the shipyard shouldn’t have to pay the full value of the stolen wood in extra wage bill to retain workers without letting them steal.

Compensating differentials can compensate for cushy, comfortable jobs, in which case, they would be negative. They can compensate for nice climate in a region, in which all jobs in a particular region would have lower wages relative to those in another region with a harsher climate. They can compensate for an expensive cost of living in a large city with high housing prices (an example of a “nontradable good”), in which case the differential is positive, relative to rural or smaller-city living. They can compensate for danger in a particular job. They can compensate for fluctuating employment in some occupations or even at the level of a particular firm, for its bad reputation for treating its employees, as we noted in subsection 10.5.3. In these examples, the workers have little or no flexibility to choose how much of the amenity or disamenity they consume on the job, once they have chosen a particular job or location. The apparent scope for discretion in the shipyard case gets smaller once we recognize that once the shipyard sets the wage, the workers literally *have* to steal a certain volume of wood to bring their wages up to the local equilibrium level. If they steal “too much” they’d be better off than their brothers in the other shipyard; some of those brothers might decide to offer their own services there, but the shipyard allowing stealing could lower their wages to maintain equilibrium – possibly in conjunction with enforcing limits on the volume of “permissible stealing.”

It will be useful to explore the mechanics of the demand for and supply of these amenities that form the basis of compensating differentials. To reduce the scope for pure preference differences that can complicate the establishment and identification of compensating differentials for things that some people commonly like and others commonly have no taste for or actually

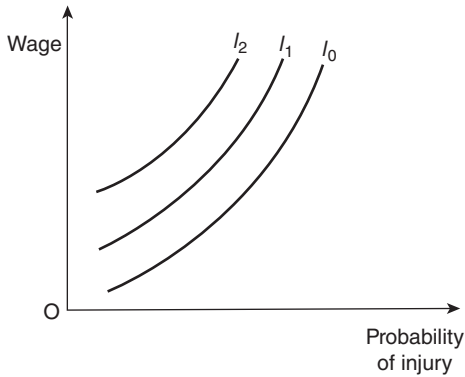


Figure 10.30 Indifference curves between wage and likelihood of injury.

dislike, we take the example of danger on the job – measured as the probability of injury. The following general exposition uses the hedonic approach, introduced in Chapter 3 as an approach to pricing differentiated goods. The full compensation package for a job is such a price for a differentiated good.

We have the same two parties to the hedonic wage transaction – the firm doing the employing and the worker. Let's suppose that the disamenity is physical danger – the probability of getting hurt on the job. Workers have a preference tradeoff between the level of their wage and the probability of getting hurt on any particular job; this comes out of the utility function. Figure 10.30 shows the worker's indifference curves between injury risk and the wage, with the level of utility increasing as we move from I_0 upward and to the left to I_2 . Along each of these curves, the worker is indifferent between the combinations of wage and risk, with a higher wage compensating for risk. The concavity of the curves indicates that the individual is risk averse: it takes larger increments of wage to compensate for equal increments to risk. If we move straight up from I_0 to I_1 , the individual is better off because she has a higher wage but the same risk. Correspondingly, if we were to move straight to the right, from I_2 to I_1 , she would be worse off because she would have the same wage but a greater risk. Figure 10.31 shows indifference curves for a more risk averse



Figure 10.31 Differences in worker's risk aversion.

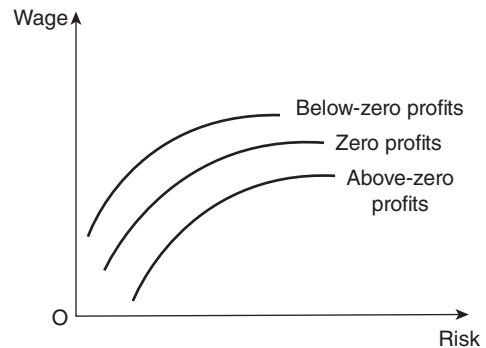


Figure 10.32 An employer's view of wage offers and risk of injury.

person and one less so. I_{more} requires a greater wage compensation for a given addition to risk than does I_{less} .

On the employer's side, the firm has to produce safety, or lower probabilities of injury, and that is costly (putting rails around vats of lye, separating operating sites so workers are less likely to hit each other with equipment, and so forth). For any given capital-labor ratio in production, a particular expenditure on safety is consistent with a fixed level of profit. Figure 10.32 shows iso-profit curves that reveal diminishing marginal returns to safety expenditures – that is, subsequent improvements in safety get smaller as the firm spends more resources on that characteristic. One of the iso-profit curves is associated with zero profits, the equilibrium condition for the

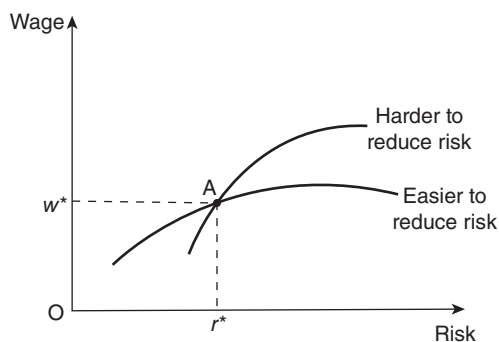


Figure 10.33 Employer's wage response when risk reduction is possible.

competitive firm. At combinations of wage rate and risk above the zero profit curve, the firm is paying too high a wage for the level of risk it is offering – not in any ethical sense of course, but just in that it can't break even if it pays wages that high yet incurs the expenditures to keep risk at such low levels. The curves in Figure 10.32 represent the technology of the wage-safety tradeoff for this particular type of firm. Other firms will face different tradeoffs, as shown in Figure 10.33. The firm facing the upper curve has to make a greater sacrifice in the wage it can offer to attract employees in order to reduce risk by a given amount than does the one facing the lower curve. Notice that at point A, where the two firms' iso-profit curves intersect, the two firms switch their order in terms of the size of the wage they can offer. At risk levels below r^* , the firm that finds it more difficult to reduce risk actually has to lower its wage below what the other firm can offer, and it still can't offer as low a risk as the other firm can. In this area to the left of point A, the lower risk-cost firm can offer more of what employees prefer while offering them less of what they don't like; it will dominate the other firm unambiguously in this region in terms of its ability to attract workers. On the other hand, the firm we've characterized as having greater difficulty in containing risk will have an advantage at risks greater than r^* , being able to offer a higher wage at any particular level of risk (or, at any wage, a lower risk).

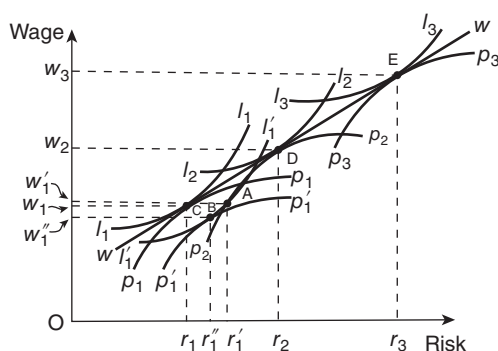


Figure 10.34 Risks and wages with different types of workers and firms.

Now, we put these two diagrams together in Figure 10.34 to show what an array of firms with different risk-reduction cost functions and individuals with different wage-risk preferences will produce in the way of an array of working conditions and wages. We show the iso-profit curves of three out of a large number of producers (hirers of labor) as p_1 , p_2 , and p_3 . Additionally, we show an above-zero-profit iso-profit curve for firm 1 as p'_1 . For workers, we show three classes of individual, each with a distinct set of wage-risk preferences represented by I_1 , I_2 , and I_3 . For class 1 we also show a lower level of indifference, I'_1 . Now we're ready to tell this story. The firms have found the combinations of wage and safety that are consistent with each of them making zero profit. If firm 1 did not exist, individuals of class 1 would find their best opportunity with firm 2, at point B, earning wage w'_1 and facing injury probability r'_1 . Individuals with preference class 2 also would be best satisfied working for firm 2 but at point D, earning a higher wage and accepting more risk, w_2 and r_2 . Whether firm 2 could segment its operations so as to be able to offer the two separate risk levels is questionable. Allowing the economy a sufficiently wide range of technologies, a firm of some type like firm 1 would find a niche in which to open business, or workers of class 1 preferences would have to accept the w_2 , r_2 combination at point D. At that point, they would not reach a tangency with firm 2's iso-profit surface but would cut

through point D instead, indicating that at that level of risk, these people would need more wage compensation to be as well off at that wage-risk combination as people of preference class 2. Thus, two different groups of people earning the same wage and facing the same risk level would not be equally well off.

But enter firm 1 with iso-profit curve p_1 , and preference-class-1 individuals would find an equilibrium at point C. If firm 1 sought to increase its profit levels above the competitive level, by lowering wages and increasing risks, as represented in surface P_1 , class-1 individuals would be able to reach only indifference level I'_1 , but on that indifference surface, they would be indifferent between working for firm 1, earning wage w'_1 and facing risk r'_1 and working for firm 2 at wage $w'_1 > w''_1$ while facing risk $r'_1 > r''_1$. Competitive pressures push firm 1's working conditions up to the wage-risk combinations represented by iso-profit surface p_1 , at which preference-class-1 workers are able to equalize their marginal rate of substitution between wage rate and risk with the firm's marginal cost of providing wage and risk, at point C, the tangency of indifference surface I_1 and iso-profit surface p_1 .

The combination of the outermost edges of the zero-iso-profit curves of all the firms is the envelope of wage-risk combinations, or the wage offer surface. Only along this offer curve can firms afford to make wage offers to potential workers, in light of the efforts they have had to undertake to create whatever conditions of safety they provide. With a large enough array of firms, the gaps between points C, D, and E will be filled in with other firms' iso-profit curves, and the line ww in Figure 10.34 will have a continuous number of wage-risk offers. Otherwise, only offers C, D, and E will exist, and workers whose tastes do not exactly match these marginal rates of substitution between wage and risk will have to make the best choice they can in light of the options. Some classes of individuals will not reach the same indifference levels that others can.

Consider a simple implication of compensating wage differentials. A worker ordinarily works 10-hour days in the sun, at the wage rate of one

small bag of dates per day. If he's willing to put in another four hours, the overseer lets him work in the shade and pays him 1.4 small bags of dates per day. His "overtime" rate exceeds his regular-time rate by an amount that we don't have the information on which to make exact calculations, since it depends on his personal valuation of working in the shade, but since he's getting the same daily rate (prorated) and gets improved working conditions, we know that his full compensation package for the last 4 hours is greater than that for the first 10.

10.6 Migration

One form of investing in human capital is to migrate. Migration is costly, and the goal generally can be construed as improving the migrant's productive capacity. Transportation requires a considerable outlay. Passage on a ship would require a fare. If the fare does not cover food and drink during the passage, these costs must be added in if the traveler is to survive. If a migrant just walks to his or her destination, walking any distance takes considerable time; food and drink again are required (somebody pays those costs, even if the migrant is given "free" room and board the entire way), possibly shelter will be desired, and the time spent walking is time the migrant can't spend cultivating fields or spinning pots or whatever it is that he knows how to do.

Considering these potentially large expenses of migrating, it seems reasonable that a potential migrant would do her best to get a good idea of what was at the intended destination that would be better than what she already has. A cost preceding migration is incurred in collecting information and assessing its accuracy to the best of one's ability. This cost may be incurred mostly in the form of time, but that time could be spent in plying one's trade (or plying it more diligently) or enjoying one's leisure. And then the dangers of traveling in antiquity were considerable and well known; they have the effect of either increasing the expected cost or reducing the expected benefit upon arrival.

Section 10.6.1 characterizes the motivations for migration and the structure of the calculations underlying those decisions. The subsequent section deals with the consequences of migration for the migrant and for the origin and destination regions. The final section turns to refugee migration, which still has many of the same structural characteristics as voluntary migration.

The population of the Italian peninsula from the second century B.C.E. (or earlier) through the imperial period experienced considerable mobility, from participation in regional fairs to seasonal migration based on agricultural cycles, to transhumance based on topographical variations combined with seasonal variations (Erdkamp 2008, 421–433). Scheidel (2004, 19–20) estimates that in the last two centuries B.C.E., 1 million to 1.25 million people were resettled in colonies or on *viridane* allotments (Rome's method of distributing newly conquered land and populating it with politically reliable settlers), for most, largely not of their own choice. A similar number moved from the Italian countryside to Rome. And on the subject of migration to Rome, the matters of hygiene, disease and mortality are prominent, raising the issue of where Rome's population growth came from – urban dwellers or immigrants. Scheidel (2003) has assembled ugly evidence on the hygiene of ancient cities, although Erdkamp (2008, 438–439) reports opinions that regular contact between urban natives and immigrants may have sent some of the urban diseases to the countryside, where people developed immunities; nonetheless, drinking fecal matter in water from an urban well or other water source would have overwhelmed the resistance of humans regardless of their origin. Erdkamp (2008, 442–444) also brings up the urban sex ratio, with a smaller proportion of women than men entering the ranks of immigrants to Rome (population growth depends on women, not on men; women are the limiting factor in population growth), but concludes that enough women moved to Rome for its population to have grown and been sustained.

10.6.1 *Economic incentives for migration*

The expected net benefit of migration is a present value – the discounted differential earnings

stream at the destination over that at the origin, net of moving costs. We can write that calculation as $PV = \sum_{t=j}^T [1/(1+r)^{t-j}][(E_{dt} - C_{dt}) - (E_{ot} - C_{ot})] - M$, in which the subscript t counts the age of the potential migrant in years, j is his age in the year of migration, d indicates a variable at the destination and o one at the origin, and the year T at the top of the summation sign is the migrant's age in the last year of his calculation horizon.²³ Within the parentheses, E_i is earnings at the destination or origin, and C_i represents costs at the same location. The discount rate is represented by r , and at the end of the expression, M represents the movement costs from the origin to the destination.

This formulation contains a lot of information. Let's start from the left-hand end of it. The term in brackets following the summation sign is the familiar discount factor – we divide by one plus the discount rate, with that sum multiplied by itself for each succeeding period. Think about the temporal structure of the decision. Suppose that T is roughly 50 years for potential migrants in antiquity: considering mortality and morbidity, 50 years might be an optimistic working life. Now drop down to the bottom of the summation sign and notice that the date of migration is the year $t = j$. The closer j is to 50, the fewer years there will be to earn a greater amount at a more lucrative destination. Jumping now to the right-hand end of the expression where we find the fixed amount of the movement cost, M , incurred in the year of migration (we could suppose that there is a market for loans to migrate, such as could be provided in an indenturing system at the destination, and probably arranged at the origin) we can expect migrants to be relatively young because migrants need a long enough time to let the higher net earnings at the destination repay them for their movement costs. And notice that the movement costs aren't discounted – they're incurred immediately even if they can be repaid over time. If movement costs have both a time and a "cash" (shekels, bags of grain, whatever "coin" of the realm) component, the higher the cash component, we would expect migrants to be relatively older unless they were financed by something like an indenture system.

The magnitude of the movement costs, M , can be an important deterrent, but more so between

some origin-destination pairs than others because of differential distances, differential terrain in between, and differential information quality. Greater distance, other things being the same, reduces migration flows by lowering the present value of net benefits from migrating. However, distance may be more a surrogate for information quality about destinations than a direct indicator of movement costs.

The terms $E_i - C_i$ in parentheses represent the earnings less the cost of living at each location. The difference between these benefits and costs determines the direction that migrants would tend to flow between any two locations i and j (supposing we didn't identify one of them as the origin and the other as the destination beforehand). But recall that it is the present discounted values of these differentials that determine the net benefits, so the time profile of earnings in particular over a potential migrant's working lifecycle will be especially important. An early and rapid rise in earnings at a destination, beginning from a level $E_{dj} < E_{oj}$, could be consistent with a rational benefit decision, particularly if the anticipated earnings stream at the origin was flat or declining over the lifecycle.

It is reasonable for migrants to think in terms of the likelihood that they will fare as well as they've heard it's possible to fare at some destination. For instance, there may be a distribution of possible earnings at the destination, with where one ends up in that distribution depending partly on skill and effort, but partly on luck. The earnings at the destination might be better formulated as $\pi E_{dt1} + (1 - \pi)E_{dt2}$, where subscripts 1 and 2 on earnings indicate two different earnings opportunities – say employment in the king's household cavalry and employment sweeping stables in a third-rate inn – and π is the probability the potential migrant assigns to landing employment of type 1. The expected value of earnings at the destination is the probability-weighted sum of the two earnings possibilities.²⁴

Migrants can miscalculate their chances of success at some destination, or they may simply not be lucky. Or the economy might change once they arrive, dashing their expectations. Return migration – going back home – can be an error-correction mechanism, or it could be simply one step in a lifecycle migration that

maximizes lifetime earnings by working in different locations at different stages of one's life. Returning home with accumulated wealth and the prestige of having seen foreign lands can be an attractive goal. Additionally, a series of migrations, as a set of stepping stones, might be part of an early plan or an early revision of an original plan: for example, a series of moves from smaller to larger towns and cities, with stays of several years at each location. Experience acquired in each stage can prepare the migrant for the subsequent move.

The $E_i - C_i$ terms implicitly appeal to potential migrants' information about locations they may never have been to. Previous migrants from one's home town, village, or region would have been an important source of information about opportunities at sites of potential relocation, as would people engaged in the distributive trades and transportation – traders, teamsters, river boatmen – although surely other sources of information existed. Additionally, immigrant enclaves within large, regional cities could help migrants from particular families or towns or of particular ethnicities or language groups adjust to urban life and find employment and possibly some occasional assistance during periods of unemployment. The common languages would have been a kind of human capital that facilitated new immigrants' acquisition of urban human capital.²⁵ Stark (1995, Chapter 5) introduces an additional mechanism capable of raising the incomes of migrants to a city relative to those of natives which shifts the focus from the human capital of the individual migrants to characteristics of the migrants as a group: lower cost of recognizing characteristics of people with whom to engage in various exchanges, for example, via common language or other cultural signals. Thus clustering of migrants from the same geographical / cultural source would tend to elevate their wellbeing in the city.

More highly skilled people in a region with a relatively compressed distribution of earnings – that is, one in which there's not that much difference between the best-off and worst-off – would be attracted to migrate to a region with a wider earnings distribution that offered them better opportunities to benefit from their skills. Lower skilled people would find movement from a

region with a less compressed earnings distribution to one with a more compressed distribution attractive. Thus, between some pairs of regions or cities we might expect to see less skilled people migrating and movements of higher skilled people between other pairs, depending on the relative returns to skills over the lifecycle in the pairs of locations.

Empirically there is some evidence of “selective” migration – that is, the more capable individuals in a population are more likely to migrate. They may have greater confidence in their ability to negotiate a strange environment. This selectivity cuts across skill levels at the time of migration: skills can be acquired in the destination as well as in the origin, but the individual daring that increases the tendency to migrate from one’s home may not be especially highly correlated with the accidents of birth into a particular family income level.

10.6.2 Consequences of migration

The typical migrant’s earnings in his new home are lower than those of natives of comparable age. Migrants may spend their first decade acquiring destination-specific human capital appropriate to the new location. This human capital is important to their earnings: recent evidence from the United States, for whatever benchmark it may serve for ancient migrants, indicates that recent migrants fluent in the language of their new country earn about half again of what those who are not fluent earn. Continuing through the working lifecycle, within about a decade-and-a-half, the migrants’ earnings are typically about on a par with those of comparable natives. Their maximum earnings reach a little more than 10% above those of comparable natives, probably reflecting the selection bias among migrants. This time profile accords well with what lifecycle labor supply theory would lead us to expect, although the specific percentages are certainly accidents of circumstances.

One of the abiding concerns about immigrants in a region receiving substantial numbers of them is their effect on wages of natives. While it is likely that immigration in large enough

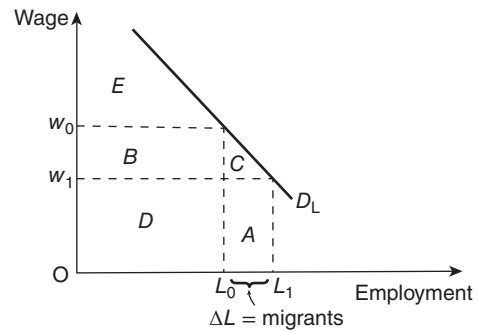


Figure 10.35 Effects of immigration on employment and wages.

volume would depress wages in the receiving region, there is more to the story. Figure 10.35 illustrates the other components. Curve D_L is the demand for labor in a country receiving migrants. Before the migration begins, there are L_0 native workers earning wage w_0 . The total earnings of labor are the sum of areas B plus D . Area E , which is the area under the demand curve for labor and above the line describing the wage rate – that is, the part of output represented by the demand-for-labor curve and not paid to labor – is the income going to owners of other factors in the region – capital and land for simplicity.²⁶ Now, enter migrants in the number ΔL , which raises the number of workers employed to L_1 . The wage rate falls along the labor demand curve to w_1 , the effect so prominently noticed. The migrants produce, and get paid, an amount equal to area A , the new wage rate times their numbers. The fall in the wage rate, accompanying a reduction in the ratio of total labor to capital and labor, induces a redistribution of income in the receiving country in the amount of area B . This much of what native workers formerly earned is now produced by other factors and the income accordingly goes to them. Don’t forget that some of the native workers whose wage has fallen may own some of these factors, so the composition of their income may have changed. Area C is an area of new production – consumer’s surplus – which goes to all residents, natives and migrants, owners of labor and owners of other factors.

The formula for the magnitude of the consumer surplus gain is $C = \frac{1}{2}(I/\eta)(\Delta L/L)^2$, where I is the aggregate wage earnings of native workers and η is the elasticity of demand for labor. The magnitude of the income transfer, B , however, is $B = (I/\eta)\Delta L/L = 2C(\Delta L/L)^2$, where C is the magnitude of the consumer surplus. Comparing what goes into making the magnitudes of areas C and B , it is clear that the income transfer, area B , will always be larger than the net welfare gain of area C . This is why income distribution concerns so frequently dominate policies toward immigration.

Figure 10.36 shows the effect of production technology on these relationships. The first thing to notice is that the D_L curve is much flatter – more elastic – than it was in Figure 10.35. Just eyeballing it, you can see that the ratio of the sum of areas B plus D to area E is larger in Figure 10.36 than in Figure 10.35: the cost share of labor in production is larger. Connect this to the greater elasticity of the labor demand curve (the marginal product of labor curve) and relate this connection back to Marshall's law of derived demand that says that the demand for labor will be more elastic the larger the cost share of labor in production. This is that law in action. Under these circumstances, the wage will fall by a smaller percentage, and the income transfer from labor to other factors will be proportionally smaller than when labor's cost share was larger. In highly labor-intensive economies, the income transfers caused by immigration will be relatively

small, and the immigration may be innocuous from the perspective of public opinion.²⁷

The consequences of migration on the receiving region also depend on what migrants bring with them. If they bring no capital, as Figures 10.35 and 10.36 assumed, the aggregate capital / labor ratio falls. If migrants are particularly well off and bring with them more capital than the average native possesses, they will raise the aggregate capital / labor ratio and may increase the demand for skilled labor complementary to that capital. Changes in the average labor income and in the aggregate composition of income in the receiving region may alter the composition of demand, although with a relatively restricted array of consumer products in most ancient societies, there may have been limited scope for this effect.

Immigrants also may send remittances back to families and relatives in their own home countries. Altogether, that magnitude cannot exceed the sum of areas A plus C . Some part of area A must be devoted to keeping migrants fed, clothed and housed, so that entire area is not available for remittances. To the extent that immigrants capture part of area C , their full share of that area could be remitted out of the new country.

The effects on the region or country of origin are largely symmetric. If emigrants take nothing but their own labor with them, the capital/labor ratio in the sending country will rise, producing a corresponding rise in the wage. If land is not taken out of production as a consequence of the emigration – and there is little reason to suspect it would have been except in the cases of massive population movements – the rent on land would tend to fall relative to the wage. Thus, the income transfers experienced in the receiving country are paralleled largely in reverse in the sending country. Emigrants in some circumstances could have sent remittances back home that were larger than their own contributions to local output. The effect of those remittances would have depended on their form – precious metals, goods, slaves. The receipt of precious metals would have precipitated some kind of “transfer” mechanism. Briefly, people can't eat precious metals, so using them to

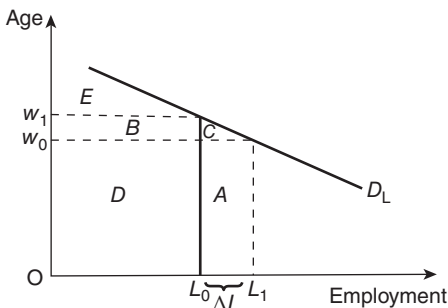


Figure 10.36 Effect of production technology on immigration's wage and employment impacts.

purchase locally produced goods can only raise the price of those goods unless additional local factors of production can be brought into use. If local factors are in fixed supply, those remitted metals must be exported to increase total real consumption in the receiving economy. Otherwise they would just raise the local price level and rearrange the personal composition of the local income distribution: remittance recipients would rise in the personal income distribution and those not receiving remittances would sink a bit.

Another important form of interaction between the receiving and sending regions is the transmission of information on the destination back to the origin. This is a prime reason why certain regions become traditional sources of migrants to other regions. The existence of earlier migrants in a destination can reduce the uncertainty about earnings possibilities there and reduce the number of migration errors. They also can facilitate the rate of accumulation of location-specific human capital and assist in the financing of movement costs and living costs while looking for employment in the new home.

10.6.3 *Refugee migration*

Both natural catastrophes and warfare have propelled waves of emigrants out of regions. Deportations of populations after conquests were common in the ancient Near East from at least as early as the Bronze Age Hittites, and the Neo-Assyrians applied the practice frequently, leaving good records (Bondi 1999, 43–44; Robertson 2005, 223–224). Underlying these migrations are net earnings calculations identical to those determining voluntary migration. Local earnings may go to zero and local costs effectively to infinity, both very quickly and with little warning. The effective alternative to emigration is death, through either starvation or murder. Not much of a choice really, but the structural equivalence gives the analyst considerable continuity in analyzing the consequences of refugee migrations.

Keeping in mind the structural equivalence of the comparative present value formulation in

voluntary and refugee migration, let's look at the differences between the two types of emigration. First, political refugees – but not necessarily refugees from natural catastrophes – generally can't go back home, regardless how difficult things may get in the new home. Consequently they can be expected to supply greater effort and diligence than voluntary migrants in similar circumstances. Second, they have less time to study alternative destinations, so they can be expected to at least start off in destinations that are economically less appropriate to their skill mix than would voluntary migrants with time to study their alternatives. Third, refugee migrants of either natural or political motivation are unlikely to bring with them any substantial amount of capital. When they arrive at a destination, capital accumulation will rank high among their economic goals. This motivation might be observable in archaeological evidence, with not much more imagination than has been exercised in envisioning immigrants in jewelry, ceramic, and other artistic and craft remains.

10.6.4 *Equilibrating migration flows when the wage rate doesn't adjust*

At certain times and places in antiquity, rural-to-urban migration was an important phenomenon, Early Imperial Rome being a case in point. The most straightforward approach to regulating migration among regions is for labor remuneration (allowing for cost-of-living differences, differences in amenities, and related factors) – roughly the wage rate – to adjust to make potential migrants indifferent between staying where they are and moving. However, in Early Imperial Rome, unemployment in Rome appears to have coincided with continuing migration inflows (Brunt 1980; Grey and Parkin 2003). Apparently (or perhaps “possibly” is a better qualifier) the urban and rural wages did not move sufficiently to eliminate the unemployment, but the question remains why migration would have continued. Push from conditions in the countryside has been suggested as a possibility, but this can't account for the failure of the wage (remuneration) differential to adjust so as to cut off the

migration. Scheidel (2004, 14) suggests that the introduction of food subsidies in the first-century B.C.E. probably accelerated migration to Rome. Alternatively, opportunities in the city, relative to those in the countryside, may have constituted a powerful attractive force but still such an explanation leaves open the same questions about the wage differential and the unemployment in the city.

The Harris–Todaro model of migration and unemployment in twentieth century Third World cities offers some insights, but should be used with care.²⁸ Roughly speaking, the model uses urban unemployment rather than the wage differential as the equilibration mechanism for migration flows, letting the nonstochastic rural remuneration equal the expected urban remuneration in equilibrium, where the expected urban remuneration is, again roughly speaking, the probability of getting employed at the urban wage times the urban wage.

The earlier Todaro model specified the rate of migration from countryside to city as a function of the percentage difference in the present discounted values (pdv) of expected lifetime earnings in rural and urban settings: $\dot{M} = \frac{A(t)}{C(t)} M \left[\frac{E_u(t) - E_r(t)}{E_r(t)} \right]$, $M' > 0$, where $\dot{M}$ is the rate of immigration as a fraction of the urban labor force, $A(t)$ and $C(t)$ are total agricultural and city labor forces, and $E_u(t)$ and $E_r(t)$ are the pdvs of expected urban and rural earnings, defined as $E_r(0) = \sum_{t=0}^T Y_r(t)/(1+i)^t$ and $E_u(0) = \sum_{t=0}^T p(t)Y_u(t)/(1+i)^t - K(0)$, where $K(0)$ is the cost of moving from countryside to city, i is the discount rate, and $p(t)$ is the probability of having a job in the city. In this framework, the urban-rural earnings differential can be positive while the expected differential is negative.

The probability of having a job in the city, $p(t)$, is defined as the cumulative probability of getting a job in each period in which the migrant is in the city, $\pi(t)$. Thus, when a migrant arrives in the city (time period zero), his or her probability of having a job there is the same as the probability of getting a job in the city during the initial period (say a quarter or a year): $p(0) = \pi(0)$; in

the next time period, the probability of having a job is the probability of having found a job in the initial period plus, if that didn't occur, the probability of getting a job in the next period: $p(1) = \pi(0) + [1 - \pi(0)]\pi(1)$. In general, over time, the probability of having a job at any date is $p(t) = \pi(0) + \sum_{j=1}^t \pi(j) \prod_{k=0}^{j-1} [1 - \pi(k)]$, in which the symbol $\prod_{k=0}^{j-1} x_k = x_0 \cdot x_1 \cdots x_{j-2} \cdot x_{j-1}$.

Next a definition of the probability of getting employed in a given period must be established. Job creation in the city is demand driven, with growth in urban employment propelled by an increase in the growth of urban output, but dampened by any growth in labor productivity: $N(t) = N(0) \prod_{t=1}^T (1 + \lambda - \rho)$. Letting the rate of job creation be $\gamma = \lambda - \rho$, the probability of getting a job in any period t is $\gamma N(t)/[C(t) - N(t)]$, where the total labor force $C(t) = N(t) + U(t)$, which says that the labor force is the sum of the people with jobs plus all the migrants who have arrived in the city but have not yet found jobs, $U(t)$. Allowing for natural population increase among both job holders and job seekers, β , and with some further manipulations and simplifications, yields an equilibrium proportion of the total urban labor force that is unemployed as $U^* = 1 - \{(\gamma - \beta)/[\gamma(1 + M^*(\alpha) - \beta) + \gamma - \beta]\}$, in which the M^* function modifies the original lifetime earnings differential to a current earnings differential: $\alpha = Y_u(t) - Y_r(t)/Y_t(t)$.

In the Todaro model, the inflexibility of wages is obvious but implicit. The Harris–Todaro model proper specifies a fixed minimum wage in the urban sector of an economy. In the abbreviated presentation of the Todaro model above, I have simplified the sectoral structure of that model to employment in the city and unemployment in the city, whereas the original Todaro model specifies an urban traditional sector, which includes both unskilled work and genuine unemployment, and the Harris–Todaro model adds a fully specified rural-agricultural sector as well. These sectors reflect the interests of development economists in the 1960s, when the economies of many Third World cities were sharply divided between modern-sector production demanding more highly skilled labor and a traditional

sector using old, sometimes ancient, production methods and often including vast swaths of underemployment – coat hanger hawkers, parked car watchers, and such. So the sectoral structure of this model in which the urban unemployment rate rather than the wage rate equilibrates migration is attuned to recent economic conditions that are not entirely applicable, at least not without modification, to ancient urban economies.

While some high skills existed in ancient cities, much of the production in cities could be picked up by recent immigrants from the countryside quite quickly, so there should not have been a major nonprice barrier to absorption into urban employment. The Harris–Todaro version of this model actually specifies a fixed (unionized) wage in the urban modern sector, which does not appear to have a clear counterpart in antiquity, even though it seems, at least in Early Imperial Rome, that the wage might not have cleared the labor market (more study of that labor market could pay dividends). A further problem with this class of model that relies on an expected wage to clear the labor market of cities (and inter-regionally as well) is that if the urban modern sector requires skills that recent rural immigrants do not have, their prospects of getting jobs other than janitorial in that sector are extremely low to zero, rendering the expected wage as a problematic market clearing device. So, while this type of model may have some useful applications to situations of ancient rural-to-urban migration and urban unemployment, it begs the issue of why the wage structure did not adjust to leave unemployment at what could be characterized as a natural, or frictional, rate of unemployment. The model may be useful nonetheless for its focus on the expectations of longer-term earnings differentials, movement costs and an uptake duration for absorption into urban employment, as well as the demand-pull effect of urban growth, as influences on a particular type of migration.

of most personal consumption decisions, many investment decisions, and reproduction decisions. Where markets are weak, extended families provide useful transaction systems. The composition of families is variable. It may include only the nuclear family members common in many Western industrial countries today – husband, wife, and children, with the occasional presence of an elderly, widowed parent. Even in more complex households and extended families occupying the same dwelling, the spousal pair may still retain reproduction decisions. In this section we will focus primarily on the nuclear family, although that focus widens a bit when we examine polygynous marriage. The approach will be frankly economic, concentrating on the influences of costs, production opportunities, and benefits rather than social mores and ethical and religious doctrines. The result may appear jarring at first to readers unaccustomed to analyzing the operation of these institutions in terms of resource allocation decisions. I hope that some of our previous application of the household production model to topics not traditionally considered economic may have prepared readers for the subsequent extensions.

The concept of the public good, introduced in Chapter 6, is useful in thinking about the family in resource allocation terms. The “family public good” is a major source of the benefits of marriage, as its simultaneous consumption by multiple family members reduces its cost to all, while individual consumers outside the family setting must provide them individually and generally at higher cost. Such goods include housing, cooking, heating and lighting, shared transportation, even children – each parent can “consume” their enjoyment of their children without reducing the enjoyment the other consumes; in an extended family setting, grandparents’ enjoyment of grandchildren does not detract from parents’ enjoyment of the same individuals as children.

10.7 Families

The family, in whatever form it may take, is an important institution in any society. It is the site

10.7.1 *Marriage*

We’ll address four topics in the economics of marriage. First are the gains from marriage,

based on production technology, the availability of markets, and informational advantages. Next we move to production within marriages. We follow two threads here: allocational implications of production technologies and personal preferences, and the economics of monogamy and polygyny. The third topic is “assortative mating,” or “Who marries whom?” How do marital partners choose productive characteristics in each other? Last, we look briefly at the search process for marriage partners.

We characterize the economic reason for marriage quite simply: a person can produce more income (the Z goods of household production theory) married than single. Let a man and a woman compare what each could produce while being single (call these amounts Z_{sm} and Z_{sf}) with what they each could produce by being married to the other (call their married incomes Z^m and Z^f). If $Z^m > Z_{sm}$, the man will be willing to marry, and if $Z^f > Z_{sf}$, so will the woman; if both inequalities are true, they will be willing to marry each other. We are assuming away a lot of incidentals at this point, such as individual preferences and the finer details of individual productive talents; we’ll introduce those later, but right now we assume that all males are identical and all females are also, the point to establish here being the gains from marriage over the single state. Now, what can cause $Z^f > Z_{si}$?

Possibly most straightforwardly, the division of labor permitted by being married lets each individual exploit his and her comparative advantage in production. Related to this source of productivity differential between married and single states is the opportunity to achieve increasing returns to scale in various household production activities. For example, producing meals for, say, two people requires doubling the “purchased” inputs (the x_i of the household production model, even though in largely nonmarket economies they may be produced by the same household with time spent outside the home but in fields – or traded for something else with the neighbor) but takes the same amount of time as cooking for one person would take. Hence a scale economy in cooking for two.

Sharing collective consumption goods is another source of productivity gain: if two people

can use the same amount of light as one person can, two people lighting up the night for themselves individually would require twice as many inputs as two people lighting up the night together. There is considerable scope for joint consumption in housing, which is a major consumer expenditure item.

There are a number of informational sources of productivity gain in marriage. Many transaction costs are obviated by routine or informal bargaining. The costs of monitoring the other’s efforts in various activities is lowered relative to the same observations conducted across family boundaries. These informational gains permit individuals within the same family to extend credit to one another with far less risk of moral hazard and adverse selection: family members know one another’s capabilities and inclinations. The same informational economies and joint incentives combine to make risk pooling within families more advantageous than between families.²⁹ What can be called “marital capital,” a combination of shared experiences and personal information, is comparable to firm-specific human capital in labor markets (or more generally in work outside the home, even if markets are weak). The marital capital specific to a particular marriage would be worthless should the marriage dissolve. Other household human capital – such as knowledge of various aspects of operating a household – would retain its value across marriages.

We look somewhat more closely now at production within marriage. Complementarity between men’s and women’s inputs in family production is one of the sources of gains from marriage, but the degree of complementarity may change systematically over the lifecycle and consequently influence the age at marriage for the different sexes. Women’s reproductive capital matures early, and their childhood experiences and household production capital may be more specialized toward the production of children than men’s, which may have longer investment periods. Combined with time required to accumulate physical capital to raise men’s productivity in agricultural activities, the lifecycle factor complementarity profiles would

tend to put men into marriages at later ages than women.

Much of our subsequent analysis will abstract from individual differences among people, commonly considered a key feature of marriage, so we discuss some routes by which such characteristics operate on marital production (Becker 1991, 122–124). First, individuals typically differ in their preferences for Z goods. Similar preferences between spouses can reduce household production costs under several circumstances: when there are scale economies, when specialized consumption capital lowers production costs of specific commodities, and in cases of joint consumption. Decreasing returns to scale in household production raises the incentives of people with different preference patterns to marry each other. The extensiveness of joint production within the typical marriage tends to reinforce the cost advantage of people with similar preferences. The individuality we tend to associate with love can be treated within the household production framework as a special case of preferences and entered into the utility function of each partner. The wellbeing of each partner can be one of the arguments in each partner's utility function. Z goods involving contact with the other spouse can be produced more cheaply when a particular partner is present in the family, and each spouse's utility can be directly entered as a consumption good in the other's utility function. Gary Becker's famous "Rotten Kid Theorem" involves the asymmetric placement of different family members' utilities in one another's utility functions – specifically, parents care about children's welfare but children care nothing about parents' welfare – with the result that an altruist (for example, parent) can induce even a selfish beneficiary (the "rotten kid") to act in ways consistent with maximizing the altruist's – indeed, the entire family's – utility (Becker 1974).

Now that we have characterized the sources of income gain possible in a marriage, we turn to an entire "market" for marriages (Becker 1991, Chapter 3). We will see how many people of each sex marry and how the gains from marriage will be divided between spouses. We begin with a

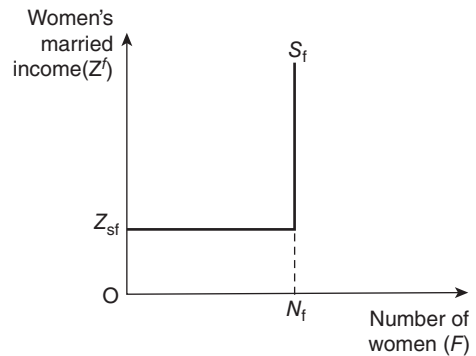


Figure 10.37 Supply curve of female marriage partners in a marriage market.

case in which only monogamous marriages are permitted, then extend the analysis to polygynous marriages. We retain for the time being the assumption that all women are identical, as are all men; we will relax that assumption later also. Figure 10.37 shows the supply curve of female marriage partners in the marriage market. There are N_f women interested in being married in this market. The supply curve of wives will be perfectly elastic (flat) when the income they can produce in a marriage is equal to what they could produce in a single household. At married income level equal to Z_{sf} on the vertical axis, the supply curve is flat, indicating that women will be indifferent between being married and single at this income. At N_f , the supply curve of wives becomes vertical, when $Z^f > Z_{sf}$: that's all the women there are, and no higher married income will bring forth any more, but if the demand for wives were to rise, they could command some more marital income (in the form of "rents").

Correspondingly, the supply curve of men willing to be married is also perfectly elastic at a married income level $Z^m = Z_{sm}$, and it becomes vertical at $M = N_m$. But, since we know "the demand for the product" (marital income, Z_{fm}), the supply curve of men is also a derived demand curve for wives (remember that the derived demand curve for a factor of production – husbands and wives can be thought of as factors of production in the production

process called marriage). Effectively, each man can offer a wife an amount of marital income equal to total marital income minus what he could have produced single and still be just indifferent between being married and being single; this amount is $Z_{fm} - Z_{sm}$. Restating this, the gains to the husband from being married must be greater than (in the limit, equal to) the income the wife obtains in the marriage: $Z_{fm} - Z_{sm} > Z_{fm} - Z^m$. The left-hand side of this inequality is the total income from the marriage less what the husband can produce being married (effectively, the amount left over to assign to the wife's production),³⁰ and the right-hand side is the difference between the total income in the marriage and what the husband could produce being single.

Consequently, the derived demand curve for wives is perfectly elastic (flat) when wives' marital income (marital marginal product) $Z^f = Z_{fm} - Z_{sm}$, and it is vertical at $F = N_m$, where Z^f is less than that amount. We add this derived demand curve for wives to the same supply curve of wives in Figure 10.38, which shows the equilibrium in the marriage market: the same number of men and women want to marry, and those would-be participants who remain single (women in this particular case) produce at least as

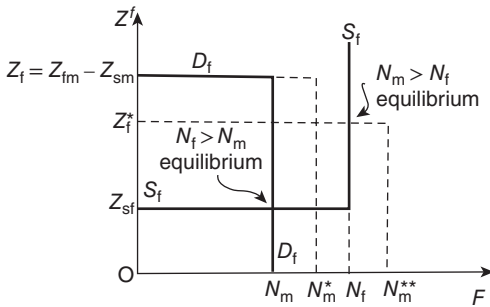


Figure 10.38 Equilibrium of women's income and number of women in a monogamous marriage market. Reprinted by permission of the publisher from *A Treatise on the Family*. Enlarged edition, by Gary S. Becker, Figure 3.1, p. 83, Cambridge, MA: Harvard University Press, © 1981, 1991, by the President and Fellows of Harvard College.

large an income as they could have by marrying. There are N_m marriages (limited by the number of men in the market), and $N_f - N_m$ women remain single. All women, single and married, produce the single level of income, Z_{sf} , while the smaller number of men receive the difference between total married output and the single income of women, $Z_{fm} - (Z_{sm} + Z_{sf})$, which is the vertical difference between the D_f curve and the S_f curve in Figure 10.38. They collect all the rent from the marriage. A small increase in the number of men (to N_m^* in Figure 10.38) would not reduce the marriage income going to men, but it would increase the number of women marrying. If the population of men increased to a number in excess of the women ($N_m^{**} > N_f$), some men would remain single, the income of all men – married and single – would fall to Z_{sm} , and the married income of women would rise to $Z_{fm} - Z_{sm}$. The reversal of the sex population size redistributes income from men to women, married and single alike. In either case, the rents available within marriage depend on single and married household production technologies.

Now, let curve S_f in Figure 10.39 represent the supply of wives to either monogamous or polygynous marriages. Their supply price will be their single income till they are all married. At that point it becomes vertical. The derived demand for first wives by the N_m identical men is perfectly elastic at $Z_{fm(1)} - Z_{sm}$ (the subscript notation "fm(1)" indicates the full marital income when husbands have one wife but could have more). However, it doesn't fall vertically to zero when each man is married, because some of them could take a second wife, who would be offered $Z^f = MP_{f(2)} = Z_{fm(2)} - Z_{fm(1)} = Z_{fm(2)} - [MP_{sm} + MP_{f(1)}]$. Let's explain the notation here. Z^f is still the symbol for wives' married income; MP indicates a marginal product, to which we referred above; $MP_{f(2)}$ is the additional output (that is, the marginal product) of a second wife, and $Z_{fm(2)}$ is the total output of a household with one man and two women; $Z_{fm(1)}$ is the output of a household with one man and one woman, MP_{sm} is the output of a single man, and $MP_{f(1)}$ is the marginal product of a first wife

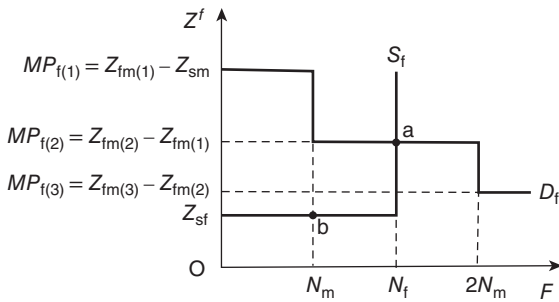


Figure 10.39 Equilibrium of women's income and number of women in a polygynous marriage market. Reprinted by permission of the publisher from *A Treatise on the Family*. Enlarged edition, by Gary S. Becker, Figure 3.2, p. 85, Cambridge, MA.: Harvard University Press, © 1981, 1991, by the President and Fellows of Harvard College.

in a marriage. In general a man with n wives would be willing to offer an additional wife an income equal to her marginal product in his household: $Z^f = MP_{f(n+1)} = Z_{fm(n+1)} - Z_{fm(n)} = Z_{fm(n+1)} - [MP_{sm} + \sum_{j=1}^n MP_{f(j)}]$. The marginal product of the $n+1$ st wife is equal to the total marital output with $n+1$ wives less the total output with n wives, which in turn is equivalent to the total marital income with $n+1$ wives less the sum of the man's single marginal product and the married marginal products of sequential wives 1 through n .

As is the case with increments to any factor of production with the quantity of the other factor held constant (the single husband in this case), there will be diminishing marginal returns to additional wives. The demand curve slopes downward in a series of steps (because of our discrete number of wives with all wives identical). Each step has horizontal length N_m (since we are assigning all men first one wife, then assigning some of them a second wife in this diagram), and the height of each step is equal to the marginal product of the n^{th} wife.

Efficiency in a polygynous marriage market doesn't require that the same numbers of men and women marry, only that the number of women who want to marry equal the demand for wives, which is what is represented at the intersection of the S_f and D_f curves in Figure 10.39. At the intersection of these two curves, all men receive married income $Z_{fm(1)} - MP_{f(2)}$ and all women receive married income $MP_{f(2)}$, whether they are monogamous or polygynous since all wives receive the marginal product of the second wife in polygynous marriages. Even though the

number of women exceeds the number of men in this case, their equilibrium married income is greater than their single income, since the women who would have been "excess" under monogamy enter into polygynous marriages as second wives rather than staying single. Women are thus better off than they would have been under monogamy restrictions, which would lower their married incomes to their single income level, less than the marginal product of second wives. Generally, if all men had at least $n-1$ wives and some had n wives, restriction to monogamy would cost each woman the difference between the marginal product of the n^{th} wife and her single income. On the other hand, the total income of men could be increased by monogamous restrictions even though total marital output would fall. In Figure 10.39, each man receives marital income $Z_{fm(1)} - MP_{f(2)}$ with polygyny (at point a), which is smaller than $Z_{fm(1)} - Z_{sf}$ (point b), which he would receive if polygyny were banned.

Now let's introduce some of the differences among people that we've held constant so far. Suppose there are two types of men (A and B), differentiated by their wealth, occupations, experience. We could appeal to the distinctions between an aristocracy and a common people or peasantry.³¹ Each group has a derived demand for wives, and their demand curves can be added together as is done in Figure 10.40. The diagram is drawn on the assumption that the marginal product of the first wives of type-B men is between the marginal products of the second and third wives of type-A men. The top-most perfectly elastic segment of the demand curve is at an income level equal to the marginal product of the

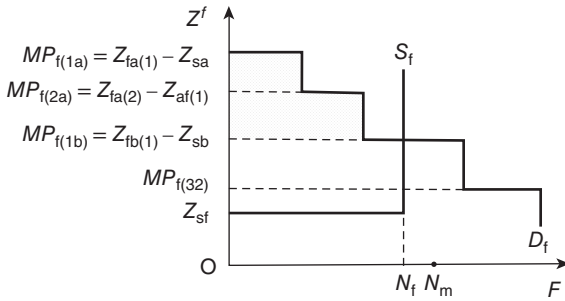


Figure 10.40 Equilibrium in a polygynous marriage market with identical women and differing men. Reprinted by permission of the publisher from *A Treatise on the Family*. Enlarged edition, by Gary S. Becker, Figure 3.3, p. 88, Cambridge, MA: Harvard University Press, © 1981, 1991, by the President and Fellows of Harvard College.

first wife in a type-A marriage, $MP_{f(1a)}$, where notation indicates the marginal product of the first wife of a type-A husband, which is equal to the total marital income of a type-A husband and a first wife, $Z_{fa(1)}$, less the single income of the type-A man, Z_{sa} . As drawn, even the second wife is more productive married to the type-A man than is the first wife to a type-B man. The third step down takes us to the income the first wife of a type-B man could produce. The last step characterizes what third wives of type-A men would produce, but there are insufficient women in the population to drive down wives' productivity to that level.

At the intersection of S_f and the aggregate demand curve for wives, D_f , type-A men take two wives each and some type-B men remain single: the total number of men of both types is $N_m > N$. All men of type A receive income $Z_{af(2)} - 2MP_{f(1b)}$, because the wives of type-B men set the ceiling on income producible by all women – the women are all the same – only the men differ – so they all produce and receive the same income at the margin. The income of type-A men includes rents in the amount of the shaded area under the demand curve in Figure 10.40. All women earn rent equal to the difference between their married and single earnings, $MP_{f(1b)} - Z_{sf}$, while type-B men earn only their single income and thus obtain no rent.

If we write down this model of polygynous marital production explicitly we can discuss how changes in some circumstances would affect the number of wives in a marriage. Let the marital production of a man of type i with number of wives w_i be a function of the total resources of the

man and the total resources of each individual wife: $Z_{m_i w_i} = w_i k(\alpha_i) Z[\rho(\alpha_i) x_m / w_i, x_f]$. The basic household production function is $Z[\blacksquare]$; α_i is an index of efficiency of men of type i ; $\rho(\blacksquare)$ is a function that converts the husband's resources x_m into an effective quantity of resources, and $k(\blacksquare)$ is another function that transforms this efficiency into different levels of output Z . The husband divides his total resources evenly among wives, so the more wives he has, the less input he applies in production with each wife, and his own marital productivity per wife falls as his number of wives increases. Each wife, on the other hand, supplies the same amount of total resources of her own, x_f , with the decreasing quantity (per wife) of husband's resources, which retards the rate of decline in wife's marginal productivity as the number of wives increases. The equilibrium number of wives increases both as husband's wealth, x_m , increases and as the husband's efficiency, $k(\alpha_i)$, increases. Since the husband employs the same amount of his resources with each wife, an increase in his wealth has the effect of increasing his demand for wives by the same percentage; the pure wealth elasticity of demand for wives is unity. The elasticity of demand for wives with respect to a change in efficiency is more complicated, depending on relative productivity properties of the husband and wives in the production function. The greater the relative marginal contribution of wives to marital output, the greater the male-efficiency elasticity of demand for wives, which could be as much as three times the magnitude of the wealth elasticity under reasonable circumstances.³²

This model gives us two sources of difference between men with more than one wife and men with one wife or none at all: wealth and productivity. The least efficient and the poorest of the men may have to remain single because they cannot compete with the more efficient and wealthier men. Of course, if we differentiated among the women as well, letting some of their productivity characteristics vary, and the wealth they bring with them as well, we would pull some of these bachelors into marriage with less efficient women. This factor will lead us just below into consideration of how people choose characteristics in their spouses, a subject called assortative mating. Before proceeding to that topic, let's consider what this model of polygyny may have to say about ancient Mediterranean societies. First, it offers to pull into the purview of ordinary economic analysis the large royal households known in nations like Egypt and city-states like Ugarit. It says that the kings may not have had so many wives simply because they could afford them but because they and the wives together could afford them. It also puts the spotlight on the household activities of those royal women as a principal locus of understanding of the sustainability of those households. We know Ramesses II had a large number of wives; why didn't he have, say, twice as many as he did? Certainly he could have "afforded" twice as many in the sense that he could have commandeered that additional number from the population of Egypt (or neighboring countries), which certainly could have found that many more willing women and their supplying families. The wives affected his own income, just as he affected theirs by bringing them within the sphere of his own resources. Splitting his time with each of them much further than he must have had to do already may have left such an increment in the number of wives less productive than they would have had to be to leave them all (each wife and the king) better off.

Returning to differences among individuals in the marriage market, we ask, "Who will marry whom?" Under what conditions will people be more likely to marry people with productive traits similar to their own (called positive

assortative mating, or positive sorting for short) or different (negative sorting)? Let's think of such traits A_i of each partner i as directly affect marital production: $\Delta Z(A_f, A_m)/\Delta A_f > 0$ and $\Delta Z(A_f, A_m)/\Delta A_m > 0$. A larger quantity or intensity of the trait on the part of each spouse contributes to marital income. A_m could be property of the men, A_f could be a particular marital skill of the woman, such as weaving. When there is positive sorting, $\Delta[\Delta Z(A_f, A_m)/\Delta A_f]/\Delta A_m > 0$, which says that an increase in the quantity of the trait the man possesses will increase the productivity of the woman's trait, and vice versa. When there is positive assortative mating and there are redundant members of one sex, as we have had in the previous diagrammatic examples, the individuals with the lowest values of these sorting traits will remain single ("undesirable bachelors"). When negative sorting dominates $\{\Delta[\Delta Z(A_f, A_m)/\Delta A_f]/\Delta A_m < 0\}$ and there is such redundancy, the individuals with the highest rankings will remain single ("desirable bachelors"). An individual's equilibrium income level and mate assignment depend on the absolute and relative levels of that person's own traits and on everyone else's. An increase in the number of men of a particular "quality" level (value of these traits) lowers the incomes of all men and raises those of women because of competition in the marriage market between men and women of different quality. With the complementarity of positive sorting, some low-quality men would be nudged out of marriage and others would be pushed into marriages with women of lower quality (trait value).

Search for a mate with the appropriate range of trait values involves a "meeting technology." The equilibrium in a marriage market (the combination of income level and the matching of spouses on these productive traits) is influenced by the costs of search and the search policies of the participants. Structurally speaking, if not necessarily institutionally, potential partners meet and compare characteristics and evaluate the gains each would make were they to marry each other. These searchers bear real costs, which increase as their intensity of search increases (the proportion of each time period they spend

searching for a marriage partner at the cost of working). The waiting time for a suitable marriage to emerge doesn't imply inefficiency in such search since this time is used to find more productive matches. The larger proportion of the population that is unattached, the more profitable will be individuals' search time, and these unattached will be more choosy about accepting partners. Absolutely larger pools also will increase the productivity of search, giving advantages to people in urban areas over rural.

Now to the institutions of search. Contemporary Westerners are accustomed to doing most of their own searching, sometimes adapting their work choices to their search strategies. Studies of the family in the ancient Mediterranean, and even much more recently than antiquity, emphasize the role of elder family members in marital search, but most of the mechanics and cost implications of the search process in the previous paragraph would carry over to such family-directed (Dad picks the bride) search. Nevertheless, when Dad picks the bride, more extensive considerations of family interrelations, such as whether a particular candidate is likely to provide drought insurance to the entire family, are likely to be included in the weighting of traits.

10.7.2 *Intrafamily resource allocation*

In our treatment of marriage we have passed rather quickly over the gains from marriage, settling for assignment of them in accordance with the usual distribution of rents among factor owners. But in the case of marriage, the factor owners literally have to live with each other, so there are continuous opportunities for negotiation. We will introduce – just introduce – several alternative concepts for thinking about how the gains from marriage could be distributed between (among) spouses. As will become apparent quickly, the greatest insights on this subject are to be derived from the analysis of strategic interactions – game theory – which gets very intricate very quickly. In this treatment of family decision making, it is useful to offer a correspondingly brief introduction of how altruism is modeled in economic analysis, and why altruism can improve

efficiency. The second topic of this subsection is parents' treatment of multiple children: how might parents choose what to give to which children? The final subject is the intrahousehold allocation of consumption and work effort, important for the insights it yields into one of the most basic locations of individual inequality – the family.

So far in this chapter, in various places, we have used the simplest characterization of family preferences, which probably isn't all that distortionary for many purposes. This formulation has collected several names: the family utility model, the common preference model, the unitary model, the common objective model, the altruistic model, and probably some others that should be recognizable as belonging to the same conceptualization: Max $U(C_w, C_h, L_w, L_h)$ subject to $C_w + C_h = w_w(1 - L_w) + w_h(1 - L_h) + Y_w + Y_h$. The family maximizes the family's utility total U by choosing the consumption and leisure of wife and husband, subject to the labor and nonlabor income each bring. In a variant on this, we have included more family members than the two spouses – children, grandparents, other relatives living in the household. Some writers have characterized this utility specification as a family social welfare function $V(U_i, U_j)$, which requires some assignment of weights to each individual's utility function, raising the issue of altruism. Another, simpler variant of the unitary model is the "dictatorial" preferences model, according to which the household head's preferences are what the family maximizes. Of course, it may be the household head who determines the weighting of utilities in the family social welfare function, and the head may be altruistic, which admits several specifications, as we see just below.

We model altruism by placing each individual's welfare within the welfare function of each other person. For instance, taking two people, a and b , if they are both altruistic (toward each other, not necessarily an array of people they don't know), their utility functions will have the form $U_a(C) = W_a[U_a(C_a), U_b(C_b)]$ and $U_b(C) = W_b[U_a(C_a), U_b(C_b)]$. If, say, individual b were selfish, his utility function would simplify to

$U_b = U_b C_b$). Altruism offers a source of intrafamily conflict, however, when we think about “merit goods” – the public goods, of which each of us thinks other people ought to have at least some minimum consumption level. If an altruistic family member enters the quantity of some such good consumed by another family member in his utility function rather than that other person’s total utility, the actions of the altruist need not be what the other family member would have him choose for him. According to altruism of this variety, $U_a = U_a(C_{1a}, C_{2a}, \dots, C_{ia}, \dots, C_{na}, C_{jb})$, while $U_b = U_b(C_{ia}, C_{1b}, C_{2b}, \dots, C_{jb}, \dots, C_{nb})$. Only part of the other person’s consumption array is in each altruistic person’s utility function, and one person’s utility weight for the other person’s consumption of good i or j need not correspond to that other person’s weighting. Alternatively, altruism can make both giving and receiving parties worse off than they would be under selfishness (no interdependence of utility functions) because of the inability to make sufficient transfers (equivalent to an altruist having two stomachs to fill) (Stark 1993, 1421; 1995, Chapter 1). Nevertheless, altruism of the general form shown above can reduce the range of disagreement within families for the simple reason that what makes one person happy also makes the other happy.

Another perspective on intrafamily – or inter-spouse – decision making can be called cooperation. This type of bargaining approach accommodates different utility functions between spouses and specifies mechanisms for reconciling preference differences.³³ We model this as one spouse maximizing a family utility function of her own subject to the joint spousal budget constraint and a constraint that the other spouse’s utility be at least a certain level: $\text{Max } U_h(C_h, C_w, L_h, L_w)$ subject to $C_h + C_w = w_h(1 - L_h) + w_w(1 - L_w) + Y_h + Y_w$ and $U_w(C_w, C_h, L_w, L_h) \geq U_w(w_w, w_h, Y_w, Y_h)$. Letting the wife’s utility depend on the wage and nonwage incomes of both spouses lets the two spouses bargain – trade control over income – to select the combination of utilities that best satisfies them. For instance, an increase in the wife’s wage income, w_w , most likely will increase her

utility U_w through two routes: it improves her opportunities outside the household (and outside the marriage) which in turn may improve her bargaining power to secure a larger share of the gains from marriage. We haven’t specified any particular bargaining mechanism for selecting the combination of utilities, but whatever mechanism they used would have to be agreed upon and enforceable.

Yet another possibility for intrafamily interaction is called noncooperation.³⁴ In fact, if spouses cannot reach agreement in cooperative strategies, they may adopt noncooperative ones, making the cooperative and noncooperative models points on a continuum of bargaining mechanisms. The “threat points” in the strategic interactions may include switching strategies, as well as divorce and even domestic violence (Tauchen *et al.* 1991). The consumption of the other spouse enters the utility function of each spouse, but each takes the other’s consumption level as outside his or her unilateral control. So each spouse attempts to solve the problem $U_i(C_i, C_j, L_i, L_j)$, subject to $C_i = w_i(1 - L_i) + Y_i$, whereby each chooses his or her own consumption and leisure levels to maximize the individual utility level, contingent on the consumption and leisure levels of the other spouse, which do affect this spouse’s utility but are outside his control, and subject to only his own income. The other spouse solves a comparable problem, but contingent on exogenous values of the first spouse’s consumption and leisure. In an alternative concept, the spouses may be responsible for provision of different household public goods and may determine their supplies unilaterally; failing to take account of the other’s demand for a public good is likely to yield less than its optimal provision level (Lundberg and Pollack 1993). The two spouses have to bargain on agreeable levels of each’s consumption and leisure, and they have more than one time period over which to bargain, renege, retaliate, and possibly reach a lasting accommodation. While people may actually solve problems of this complexity – or if they don’t actually “solve” them, they may be able to keep cycling through an acceptable set of transient allocations – problems with this large a space for strategic interactions

currently are intractable.³⁵ Analysts have to pick a much simpler set of features if they are to make predictions about how any behavior will proceed.

There is contemporary empirical evidence from both industrial and developing countries, reported by Bergstrom (1997, 39–40) that the distribution of income within households – that is, how much the wife brings in independently of the husband – affects the distribution of consumption of both spouses and children as well as the wife's supply of labor outside the household. From Thailand comes evidence that an increase in a wife's unearned income from outside the household has a larger negative effect on the probability of her working outside the household than does an equal increase in the husband's asset income. In Canada, the budget shares allocated to goods associated with men and women are sensitive to the share of family income contributed by the two spouses, and in the Ivory Coast, an increase in the proportion of household income accruing to women raises the budget share of food and lowers the shares of alcohol and cigarettes. In Brazil larger unearned income of mothers has a stronger (positive) effect on fertility and measures of child health such as calorie intake, height, weight, and survival probability, while in Thailand, an increase in a woman's unearned income increases her fertility while an increase in her husband's unearned income has no effect on it. While we still haven't introduced any specific mechanisms for making the transfers of utility between spouses, these examples of which spouse gets what consumption under different circumstances gives a round hint: one spouse can offer additional consumption to the other. This can be in the form of specific consumption or in the form of a fungible commodity like money.³⁶

We present these alternative specifications not simply to bamboozle the reader with more choices for analysis than he or she might ever want but to alert the reader to the care that must be taken in thinking about intrahousehold decision structures: even the simplest model can be formulated in quite a diversity of ways, which may produce different answers or even different questions.

As we noted, for many questions, such as labor supply response, consumption of different classes of goods, investment in training, the simplest specification – $U(C_i, L_i)$ – frequently gives useful and robust (that is, they don't change a whole lot if we "tweak" some parameter) insights, but the examples of the previous paragraph indicate that there is important scope for bargaining between spouses – and by extension, between children and parents – that the unitary model does not capture. To argue a lot more closely about the subject of intrahousehold decision making, we owe it to ourselves and our discussants and disputants to be quite precise.

The second problem we consider in this section deals with the choices parents make in preparing their children for earning their own livelihoods.³⁷ We begin by showing several ways one could conceptualize how parents assess the future wellbeing of their children vis-à-vis their own current wellbeing. A simple way of formulating children's wellbeing uses the same formulation we used in the family utility function above: $U = U[C_p, V(W_1, \dots, W_n)]$, where C_p is parents' own consumption, V is the parental welfare function, which we show as being a function of the adult wealth levels of the various children. A change in adult wealth level of any child would have the effect on the parents' utility $(\Delta U / \Delta V)(\Delta V / \Delta W_i)$: starting from the right in this expression, the change in a child's adult wealth first affects the value of the parental welfare function V , which in turn affects the parents' utility function. This method of characterizing parents' attitudes toward their children says that the parents don't care how the children get the wealth. An alternative formulation, which has proven insightful, proposes that parents care to some extent about the sources of their children's adult wealth, specifically, how much of it comes from their own earning ability, Y_i for the i^{th} child, and how much the parents give the child outright in the form of a transfer, T_i , either while the parents are living or as a bequest. This specification can be written as $U = U[C_p, V^*(Y_1, \dots, Y_n), V^{**}(T_1, \dots, T_n)]$, in which V^* is the parental subwelfare function for children's adult earnings and V^{**} is the parental

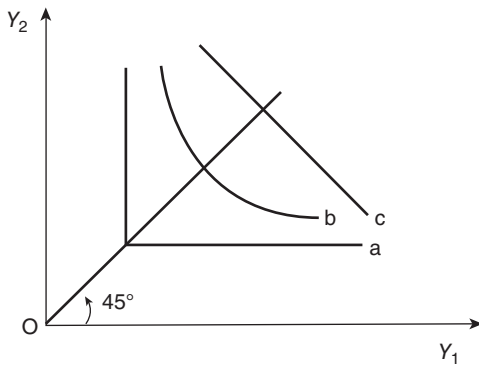


Figure 10.41 Parental indifference curves for incomes of two children.

subwelfare function for parental transfers to the children.

From this latter formulation of parents' preferences, called the separable income-transfer model (because we can separate parents' choices about influencing children's adult income and making direct transfers to them), we can draw pictures of parents' feelings (preferences) about how they should treat their children relative to one another. Figure 10.41 shows indifference curves of the parental subwelfare function for child income, for two children (showing these curves for more than two children at a time requires as many dimensions as children; three is the most that can be shown directly on paper, and it is as likely to obfuscate as enlighten, so I stick with the two-children case). The adult incomes of the two children are on the two axes, and the 45° line identifies the points of equal income between them. The L-shaped indifference curve, labeled a, shows parents concerned exclusively with equity between these two children. If one of the children were to earn a little bit more, or even quite a bit more, unless the other child's income also increased, the parents would be completely indifferent. The right angle in a family of these indifference curves moves out on the 45° line; any points off the 45° line, although they may contribute to the wellbeing of the particular child, contribute nothing to parental utility. Indifference curve b demonstrates a parental tradeoff between the incomes of the two children: some change in

the ratio of the two children's income is "acceptable" to the parents, inasmuch as different income ratios are consistent with a given level of parental utility. This acceptance of a tradeoff between the incomes of different children is equivalent to an equity-efficiency or equity-productivity tradeoff. The third indifference curve shape, labeled c, with a slope of minus one as drawn here, shows perfect substitutability between the incomes of the two children: parents with this shape of indifference curve are concerned exclusively with productivity of their children – not at all concerned with equity. The all-equity and all-productivity shapes are the limiting cases of this preference; in most cases we could expect some degree of curvature as demonstrated by curve b.

The exclusive concern with equity demonstrated by indifference curve a in Figure 10.41 is not the same thing as equal concern for children 1 and 2, as Figure 10.42 makes clear by shifting the apex of the L-shaped, equity-only indifference curve off the 45° line and indicating a parental preference for child 1. Similarly, indifference curves demonstrating a parental willingness to tradeoff income between children could be tilted so it did not have a slope equal to minus one where it cuts the 45° line, and the efficiency-only indifference curve need not have a slope of exactly minus one. The meaning of the "equity-only" shape of an indifference curve means that the distribution of income between

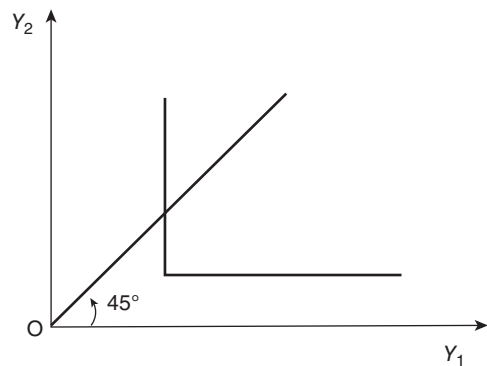


Figure 10.42 Parental willingness to trade off income between children.

children is of overriding importance to parents, not necessarily that the distribution be equal (or $1/n$ among n children).

Now we turn to how the parents can contribute to the adult earning capacities of their children. Think of a production function for adult earnings, consisting of the child's native endowments (intelligence, health, and so forth) and human capital investments made by the parents while the child is young. Such a production function is $Y_i = Y(H_i, G_i)$, where H_i is the parents' human capital investments in child i and G_i is that child's native endowments. The production function has the usual characteristics, $\Delta Y_i / \Delta H_i > 0$, $\Delta Y_i / \Delta G_i > 0$, and of particular importance, $\Delta[\Delta Y_i / \Delta H_i] / \Delta G_i > 0$, which says that human capital investments are more productive when spent on children with superior native endowments. Stated alternatively, greater native endowments permit children to transform a given level of human capital investment into greater productivity. Parents face a budget constraint on their human capital investments in their children, which we can characterize as $\sum_i p_{H_i} H_i \leq R$, in which p_{H_i} is the unit price of human capital investments in child i and R is the parents' resource constraint on their human capital investments.

If we have the parents maximize their subwelfare function for children's income subject to this production function and their resource constraint, we get a transformation curve between the adult incomes of children 1 and 2, shown in Figure 10.43. As drawn, child 1 is more easily educable than child 2 (or faces a lower, child-specific human-capital investment price), and with a symmetric indifference curve (showing equal concern) demonstrating a willingness to trade off equity for productivity, these parents would maximize their utility by investing so as to produce a higher income in child 1 than child 2. The parents still have an additional tool to affect their children's welfare – the transfers that are independent of the income the children are capable of producing themselves. Larger transfers may be made to children with lower productivity in using human capital investments.

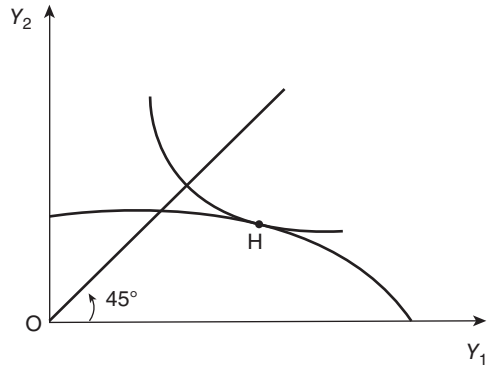


Figure 10.43 Parental production function for children's earnings.

Neither Figure 10.43 by itself nor the positive cross-productivity of H_i and G_i in the child earnings production function by itself tells whether parents will reinforce natural-endowment differences by their investment choices or compensate less well endowed children by investing more in them, thus narrowing the earnings gap between unequally endowed children. This compensation or reinforcement excludes the effects of direct transfers. Parents will compensate if the elasticity of substitution between children's incomes in the income subwelfare function is less than one and reinforce if it is greater than one. Compensating or reinforcing behavior is the "market" expression of preferences for equity versus productivity, preferences regarding the equality of concern for various children, and the technical parameters of the children's human capital production functions. As market outcomes, children's adult earnings will be more disparate the greater the relative preference of parents for productivity relative to equity, the greater is unequal concern for different children, the greater the disparity in child endowments, and the more productive are investments in human capital. Of course, some of these factors could offset each other, so we must be careful to interpret the statement above as referring to each of these factors separately. Unequal concern for harder-to-educate children could offset what is otherwise a greater preference for productivity over equity combined with unequal endowments of children.

Of what use are these concepts in the study of ancient Mediterranean societies? First, it may be thought that these ideas are difficult to examine empirically, and that is true. However, the subject is yielding answers on the topic of compensation versus reinforcement to careful statistical examination of detailed, individual and household survey data from both developed and developing countries, of the contemporary period of course. Now, for its application to the ancient periods, which are the primary interest of readers of this text, the model offers the concepts with which to interpret principally textual evidence of the ancient treatment of children: where to look for evidence – training dolts as well as geniuses, investing in sons and making transfers (for example, dowries) to daughters; and how to interpret what is there – for example, occupational choices of younger sons, with *inter vivo* financial assistance from parents, when primogeniture laws govern inheritance (compensation through differential training and bequests). Even some archaeological evidence, such as the recent discoveries of the tomb containing numerous children of Ramesses II (Weeks 2000), may be amenable to such questioning.

Finally, I turn to the subject of intrahousehold allocation of consumption and effort. Recall that what we call effort is basically labor supply and that labor supply choices are determined in the same behavioral nexus that determines consumption: the two sets of choices are not independent, even if a third party, such as a household head, is making or heavily influencing the choices. To illustrate these intrahousehold interactions, I present the highlights of a study of the interactions among consumption, health, and effort in rural Bangladesh from the early 1980s (Pitt *et al.* 1990).³⁸ I believe the social and institutional circumstances have important structural similarities to those in the ancient Mediterranean basin – largely agricultural work, some off-farm work, predominantly family enterprises, some restrictions on women's work, much of what is consumed is self-produced.

The analysis studies a family's allocation among its members of food (it's possible to include other consumption goods too, but I omit these to simplify; not a lot is lost in the omission),

work effort, and health. The family has a set of preferences regarding how these items are allocated, described in a family welfare, or utility, function. In addition to a family budget constraint on goods (effectively, cash) and time, the family has a production function for each member's health, and health in turn affects the productivity of each family member's work effort. The family welfare function should be familiar by now: $U = U(h_1^k, \dots, h_{n_k}^k, c_1^k, \dots, c_{n_k}^k, e_1^k, \dots, e_{n_k}^k)$, in which the superscript k indicates a particular class of family member, for example, male children, female children, adult females, and so forth. Utility is a function of each individual's health, h_i^k , where the subscript i indicates individual i in class k (we number the individuals in each class from 1 to n_k , the last of which means "individual n in class k "); each individual's food consumption, c_i^k , and each individual's work effort, e_i^k . The family members can directly choose only food consumption and effort, but both of those items have a direct influence on utility and an indirect influence through its effect on the individual's health.

The health production function, which is specific to each class of persons, is $h_i^k = h^k(c_i, e_i, g_i)$, where g_i is the individual's endowment of health.³⁹ According to this health function, health increases as food consumption (properly converted to nutrition content) increases, as the level of endowment (a stock measure, not a flow) increases, and as the level of effort decreases; more effort "uses up" health. In our common notation, $\Delta h_i^k / \Delta c_i > 0$, $\Delta h_i^k / \Delta g_i > 0$, and $\Delta h_i^k / \Delta e_i < 0$. The wage rate of each individual i in each class k is affected by both the effort one puts forth and one's health status: $w_i^k = w^k(e_i, h_i)$, where $\Delta w_i^k / \Delta e_i > 0$, $\Delta w_i^k / \Delta h_i > 0$, and the interaction between effort and health has a reinforcing effect on productivity: $\Delta[\Delta w_i^k / \Delta e_i] / \Delta h_i > 0$.

The family allocates consumption and work effort so as to maximize the welfare function above, subject to these two production functions (health and work productivity), and subject also to the budget constraint $\sum_k \sum_i w_i^k - p \sum_k \sum_i c_i^k + v = 0$ where p is the price of food and v is asset income. To "solve" this family allocation problem and see what its theoretical predictions are, we take the first-order

conditions for food consumption and work effort and set them equal to zero. We can get some initial information from studying the expressions resulting from that set of operations. Next, we can vary the exogenous health endowments of particular types of family members to see how an individual's consumption, work effort, and health depend on his or her own health endowment, as well as how that same individual's consumption, effort and health depend on other family members' health endowments. The responses of each of these endogenous variables to endowments are indicators of compensation for or reinforcement of an individual's health endowment.

Recall how, from first-order conditions on utility or profit maximization problems, equilibrium allocations are characterized by the equality of ratios of either marginal utilities or marginal costs to commodity or factor prices: $(\Delta U/\Delta c_i)/(\Delta U/\Delta c_j) = p_i/p_j$ (or, in the simpler notation, $U_i/U_j = p_i/p_j$) for utility maximization, $F_L/F_K = w/r$ for labor and capital in profit maximization (or cost minimization equivalently). Using this same procedure of equating ratios of marginal utilities to marginal costs (consumption costs or prices) in studying allocations of consumption across families, we get $(U_c^j + U_h^j h_c^j)/(U_c^k + U_h^k h_c^k) = (p - w^j h_c^j w_h^j)/(p - w^k h_c^k w_h^k)$. The allocation of food depends on how the household values the health and consumption of each family member (the U_h^i and U_c^i and similarly for members of class j), how the relationship between health and consumption in the health technology (the h_c^i) and between health and labor productivity (the w_h^i) differ among family members, and how the returns to investments in the health of individual family members (the wages w^j) differ. A household will tend to distribute more resources to members with higher earnings capabilities when consumption by different individuals are substitutes in the household welfare function. The marginal cost of allocating food to an individual is lower the more responsive is that individual's health to more nutrition and the more responsive her productivity (wage) is to her health. If different classes of family members participate in different work (say, by sex or age) and the productivity effects of health vary across those types of work, the marginal cost of food

allocated to different classes of family member may vary considerably. Within a given class of family member, the distribution of food and work effort across individuals would still depend on the distribution of health endowments.

In terms of consumption, the Bangladesh study found reinforcement of endowments on balance, with a much stronger effect for males than for females. Reinforcement was substantial for males age 12 years and older and for females aged 6 to 12 years, and compensation may have occurred for children of both sexes under the age of 6. Adult males with greater health endowments were more likely to undertake particularly energy-intensive work, while adult female health endowments were relatively unimportant, compared to males, in determining their activity choices. However, using another measure of reinforcement or compensation, the elasticities of own health with respect to own health endowments were 0.88 for adult males and 0.97 for adult females, implying that these households exhibited net compensatory behavior with respect to adult health endowments, "taxing" adult males, more than females, to benefit other household members.

In South India, during the agricultural lean season, when food is relatively scarce, the parental equity-productivity tradeoff is close to a preference for pure productivity, with unequal concern favoring older children over younger and sons over daughters. However, during the surplus season there is significant aversion to inequality and much more equal concern. Altogether, younger children and daughters are more vulnerable to nutritional risk during the agricultural lean season, and there is evidence that favorable weather shocks increase the probability of daughters surviving to school age (relative to the same probability for sons) (Behrman 1988a, b; 1997).

For cross-effects between the health of some family members and the activities of others, we find information from 1980 Indonesian survey data. Teenage daughters were significantly more likely to increase their participation in household child care activities and reduce their participation in work outside the household, relative to sons, in response to increased mortality of infant siblings (Pitt and Rosenzweig 1990).

10.7.3 *Children and the economics of fertility and child mortality*

It takes time to raise children, food to feed them, produced or purchased goods to clothe them, and a host of resources to invest in whatever types of human capital they will possess. With the links to the alternative uses of all these resources, it seems reasonable to study several aspects of children as a resource allocation problem. One of the principal questions to be addressed is how will the number of children a couple has respond to differences between couples in income and changes in a society's income over time. Let's think of children as "goods" for a moment, as contrasted to "bads." Children are one of the great sources of joy that adults experience, so it seems reasonable to think of children as entering a couple's utility function. As far as goods are concerned, there's no reason to think that children would be an inferior good – one that people would want to consume less of as their income increases – which leaves an interpretation of them as normal goods reasonable. Yet one of the salient empirical regularities in countries that have passed the "Demographic Transition" – in societies that have not yet entered the rapid-population-growth period of the transition between those two states – is the negative association between parents' income and number of children. Exceptions can, of course, be found, and the tendency may have been somewhat attenuated in the ancient societies of the Mediterranean Basin. Such exceptions need to be explainable in terms of the same sets of forces that could explain the purely post-Transition regularities.

We've offered some round hints as to how we will proceed with the analysis of the resource allocations involved in having and raising children. Begin with a couple's utility function with children and all other consumption as arguments:⁴⁰ $U = U(C, Z)$, where C is children – sometimes called "child services." There are two principal routes by which increasing parental income could depress the number of children in a family: children cost more to high-income families because those parents' time is more valuable; and a tendency of higher-income parents to invest

more in raising each child, which has the same effect of making children cost more per child to such parents. The first effect can be incorporated into the analysis through the parents' wage rates in the budget-cum-time constraint. The second effect needs a distinct mechanism for its operation. We can specify a production function for child services (or just children, for short) that introduces the concept of "child quality," or the household investment in each child: $C = NQ = q(t_c, x_c)N$, where t_c is the parental time devoted to raising a child, x_c is the produced or purchased goods devoted to children, N is the number of children, and $Q = q(\blacksquare)$ is quality per child. We substitute this formulation for the production of children into the utility function and maximize utility subject to the full-income constraint $I = \pi_c NQ + p_N N + p_Q Q + \pi_z Z$, in which the π_i are the cost-minimizing shadow prices of Z -goods and the household portion of child services, p_N is the outside ("market") price of the part of total child costs that is independent of child quality, and p_Q is the price of that part of total child costs independent of number of children. If there are no such independent Q - and N -specific costs, these two terms will drop out of the budget constraint. This budget constraint, in contrast to the budget constraints we've used so far, is nonlinear because of the term with the joint effect of N and Q . We can't simply add up the number of children and multiply them by a price per child and do the same for quality. The income elasticities of demand for N , Q , and Z must satisfy the relationship that $\alpha(\epsilon_Q + \epsilon_N) + (1 - \alpha)\epsilon_z = 1$, where α is the share of family income devoted to children and the ϵ_i are income elasticities. If children are normal goods, $\epsilon_Q + \epsilon_N > 1$, but it is still possible that $\epsilon_N < 0$ if ϵ_Q is large enough.

Figure 10.44 shows the nonlinear budget constraint in $Q - N$ space, and indifference curves tangent to them. As drawn, the increase in family income represented by the movement from budget line C_0 to C_1 brings forth a barely perceptible increase in the number of children; the figure easily could have been drawn to show a decrease. The indifference curves must have more curvature than the budget constraint does or

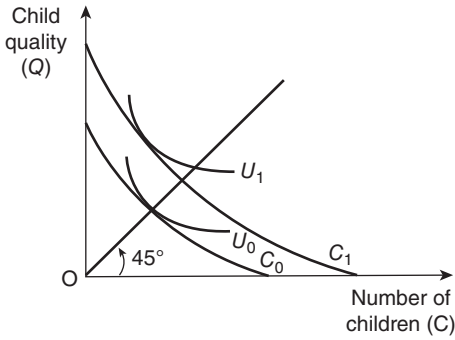


Figure 10.44 Tradeoff between child quality (investment per child) and number of children.

tangencies would represent utility minimization instead of maximization. This consideration in turn implies that the substitutability between quantity and quality cannot be too great. It is the nonlinearity of the budget constraint that causes the quantity-quality interaction that induces the substitution effect against the number of children and in favor of quality as income rises if $\epsilon_Q > \epsilon_N$. To see this clearly, from the budget constraint, first notice that the marginal costs of Q and N are $MC_Q = p_Q + \pi_c N$ and $MC_N = p_N + \pi_c Q$. Next, recall that the first-order condition for utility maximization, requires that the ratio of marginal utilities equal the ratio of marginal costs: $MU_N/MU_Q = MC_N/MC_Q = Q/N$. Thus the relative cost of the number of children tends to increase as Q/N increases, which will happen if $\epsilon_Q > \epsilon_N$.

The other principal reason for the negative relationship between income and fertility is that higher income is associated with a higher cost of female time, either because of a higher female wage or because a higher household income (either from the husband's labor income or from asset income) raises the value of female time in household activities. If raising children is relatively time-intensive, particularly for others, the opportunity cost of children tends to rise relative to the costs of other sources of satisfaction, which induces substitution against children as a consumption good. Figures 10.45 and 10.46 show the effect of increases in a wife's and husband's

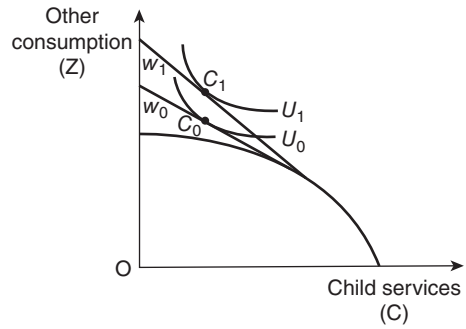


Figure 10.45 Effect of wife's wage on fertility.

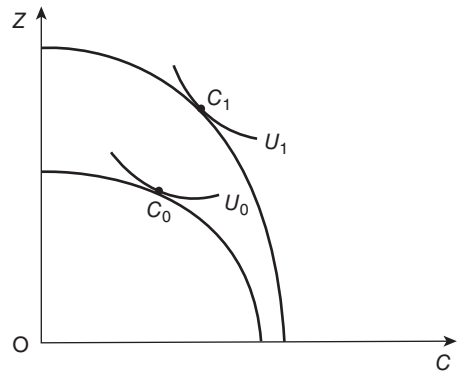


Figure 10.46 Effect of husband's wage on fertility.

labor income on fertility. Both diagrams show a transformation curve between child services, C , and other consumption, Z . In Figure 10.45, the wife's wage rises from w_0 to w_1 , and the tangency with the couple's indifference curve U_1 yields child services $C_1 < C_0$. But $C = NQ$, so do both N and Q fall or does one fall and the other rise? Intuition suggests that it would be unusual for child quality to fall while parental income is rising, so there is a presumption that quality rises and number of children falls as the wife's labor time becomes more valuable. Turn to the increase in the husband's wage in Figure 10.46. Since the husband is not the principal producer of child services, the transformation curve expands as a consequence of his wage increase, and it expands more in the direction of other consumption, which can be more readily increased with his

greater earnings, than in the direction of child services, which are relatively intensive in the wife's time. Child services increase, and intuition is insufficient to suggest as strongly that number of children fall while quality per child rises. In fact, fertility may increase with increasing husband's labor income. A higher female wage can delay marriage and reduce fertility by reducing the length of married time the woman is able to bear children.

In general, high fertility rates in low-income societies occur not from ignorance or short-sightedness but because the time costs of raising children are low and the costs of regulating fertility are high. Birth control would be practiced only if a couple's demand for births exceeded its supply. When mothers' and children's productivities are about the same, the loss of a mother's productivity through pregnancy and child care can be compensated easily by children's household working time.

If parents have a target number of children, an increase in the survival rate reduces the real price of a surviving birth, which would reduce the number of births if the demand for fertility is price inelastic, but the number of surviving children must increase as the survival rate increases. Parents can practice two strategies in response to child mortality: hoarding and replacement. Hoarding is a response of fertility to expected mortality: having more births than the desired number of surviving children in anticipation of infant and child deaths. Replacement is the response to experienced child mortality. If deaths are prevalent at very young ages, replacement can serve to replenish the lost children at a feasible time in the lifecycle and may largely obviate hoarding. If families have a very rigid target for survivors, and they practice hoarding to assure their target, they are likely to overshoot and end up having to support more children than were necessary.

Many of the behaviors that are expected to affect infant and child mortality can be modeled as utility-maximizing choices by families, and at the level of the family, mortality and fertility may be determined jointly, although the resulting mortality need not be considered deliberate.

High fertility may contribute to high mortality. For example, the birth of a new child may lead to early weaning of the preceding child; poor nutrition, diarrhea, and death may ensue for the earlier child. Mortality also can contribute to fertility by a sort of reverse route: the death of a breast-feeding child stops the mother's breast feeding, which in turn makes her more likely to incur an earlier pregnancy.

10.8 Labor and the Family Enterprise

Thus far in this chapter we've built up many of the components with which to analyze the behavior of people who work for themselves instead of for someone else. We've also seen how work decisions are determined simultaneously with consumption decisions. We now possess the building blocks for studying the behavior of the most common production unit of antiquity – the family farm – in circumstances of weak or non-existent markets for some goods and services and various degrees of interference with allocation decisions. Of course at various times and places, other production institutions appear to have been prominent, such as the temple estates of both Egypt and Mesopotamia, but both of those institutions faced the problem of how to elicit the needed labor supply (unless we modern scholars are willing to ascribe a combination of remarkable powers of coercion and equally remarkable degrees of accommodation on the parts of individuals). The production teams on mainland Chinese (People's Republic) collective farms during the third quarter of the twentieth century C.E. may offer a reasonable correspondence to the supply of labor by individual families to these large, seemingly monolithic production agencies of antiquity. Subsection 10.8.3 presents the analysis of behavioral restrictions derived from a study of recent Chinese agriculture. We begin with the formal structure of what we mean by the family farm and some analysis of how such an enterprise would use labor if labor can be hired in and hired out as desired. The second subsection analyzes the interrelatedness of production and consumption

decisions when markets for some goods or services either don't exist or operate very imperfectly, and the final subsection addresses the subject of restricted transactions which, in the limit, could entail the nonexistence or dissolution of some markets.

Cahill's (2002, 236–261; 2005) studies of activities ranging from clearly domestic to industrial in the archaeological remains of houses at fourth-century B.C.E. Olynthos and sixth-century B.C.E. Sardis show a correspondence in Mediterranean / Aegean antiquity to the kinds of activities organized by the farm household model (even if these houses were in cities rather than on farms, the relationships between households and markets undoubtedly were similar); Tsakirgis does the same for industrial activity in fourth-century B.C.E. Athenian houses, bringing literary evidence in addition to archaeological; and Ault (Tsakirgis 2005, 77–81) reports fourth-century B.C.E. archaeological evidence from the small, southern Argolid walled harbor city of Halieis, which he interprets as demonstrating connections to markets outside the home and probably outside Halieis.

To give a preview of what follows, when markets for all the farm household's inputs and outputs exist (and labor can be one of its inputs and one of its outputs), the household responds to the prices determined in those markets. When even one of those markets doesn't exist, the household's behavior imputes a shadow price (sometimes called a virtual price in this literature) that reflects the scarcity of the good or service and its value in the household. In this latter event, the determination of the shadow price for the good whose market is missing may (and, in general, will) influence the allocations of all the other goods that do have market prices. Even if all markets exist but the transactions that the family can conduct in one or more of them are restricted, say through quotas, family-specific shadow prices will again arise. These complications can give rise to phenomena such as negative supply responses to prices without any implication of irrationality.⁴¹

Concomitantly, when all relevant markets exist, the farm household's production decisions can

be made without reference to the household's consumption decisions – the utility-maximizing consumption decisions don't form shadow prices that differ from market prices – although events and decisions in the realm of production do affect consumption decisions. When all markets exist and production decisions can be made independently of consumption decisions (that is, the analyst can study them as if they were made separately), “separability” is said to exist – separability between production and consumption decisions – but note that even under these conditions, consumption decisions are still dependent on production decisions and conditions. To get a hint of why the consumption decisions still depend on the production decisions in a way that can be, but generally isn't, included in the standard model of consumption behavior, recognize that external (cash) income derives at least in part from what the family produces and sells off-farm. The income effect of a price change for a good that is consumed but not produced by the consumer is negative – a higher price is equivalent to lowering the real income of the consumer, as discussed in Chapter 3. When the consumer produces this good as well as consuming part of it, there is a mixed message in product price changes. If the price goes up, the bad news is that it costs him more to consume this good, but there's offsetting good news that what he can get for what he makes has increased – a positive income effect offsetting to some extent the negative income effect coming from the influence on the opportunity cost of consumption. Let's get into the details now.

10.8.1 *The farm family household and the separability of production decisions from consumption decisions*

We introduce the family farm under simple conditions. Consider a situation in which the owner of a farm – a household head – can either hire labor to help out on his farm or can send labor hours from his family to work off the farm.⁴² That is, there is a labor market, and the productivity of family labor on the farm is the same as the

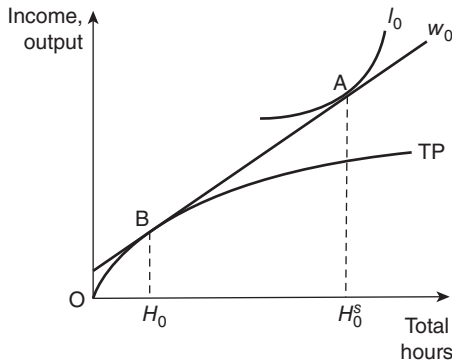


Figure 10.47 Labor supply by a family enterprise.

wage at which additional labor can be hired in or family labor hired out. In Figure 10.47, the curve TP is the farm's total product curve (total product of labor), and the wage rate is w_0 . Curve TP represents the farm family's production function. This diagram is essentially the labor supply diagram we used in section 10.3, but with the direction of working hours reversed on the horizontal axis, and the indifference curves correspondingly reversed. At wage w_0 , the household will want to supply H_0^s total hours of work, which will let the family indifference curve reach its highest tangency with the wage rate, at point A and indifference level I_0 . However, back on the farm, it only pays the family to use H_0 hours of labor on the quantity of land it has when the wage is this high. Consequently, the family will produce only at point B, where wage w_0 is tangent to the production function: marginal product of labor (the slope of the TP curve) = marginal cost of labor (the slope of wage line w_0). So the family supplies H_0 hours of work on their own land and works $H_0^s - H_0$ hours in the labor market (that is, for other people). Family labor, or that of anyone else the family might hire, is too unproductive on the family land and too valuable in the labor market for the household head to hire farm laborers or use his own family's time working his own land. The family combines some self-employment with some work for others to maximize family utility.

In Figure 10.48, the same farmer, with the same production function and quantity of land (represented by the same TP curve) faces a lower

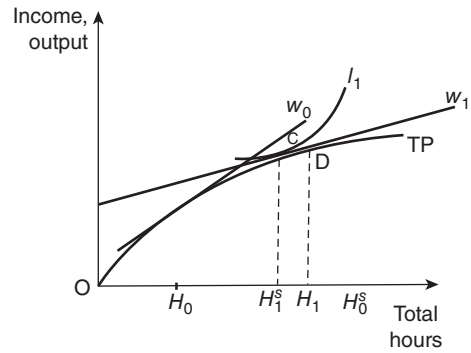


Figure 10.48 Effect of outside wage on a family enterprise's labor supply.

wage, w_1 . The family can earn less by supplying hours to the market and can earn more by working for itself. With the lower wage, the family wants to work less, and its total labor supply falls to H_1^s , where indifference curve I_1 is tangent to wage line w_1 , at point C. At the same time however, with labor so cheap, it pays to put more labor into the family farm. It is optimal to put H_1 hours into working the farm, and with the family supplying only H_1^s hours, the household head hires in $H_1 - H_1^s$ labor hours. At lower wage rates, the family offers less of its own labor and hires in more labor. This is a useful first view of the production side of the family farm, but you'll notice that we've skirted entirely the consumption side of the farm's life. We turn now to a more formal, but more complete, view of the entire problem.

We can characterize the family farm in terms of a utility function, a cash budget constraint, a full-income constraint that incorporates time constraints, and a production function that describes the farm's technology.⁴³ The "production function" can be thought of as a set of production functions for the household's entire array of activities, from ploughing fields to preparing meals. We can show the relationships we want to study here by positing just three goods: a good purchased "on the market" by the household but otherwise not produced by it, consumed in the amount X_m ; a good produced by the household, Q_a , part of which it may consume, X_a , the remainder of which it sells outside the household,

$Q_a - X_a$, in an amount commonly called “marketed surplus”; and another crop that it does not consume itself but sells outside the farm, Q_c . The utility function is $U = U(X_m, X_a, X_L)$, where X_L is leisure. The budget constraint is $Y = \sum_{i=m,a,L} p_i X_i$, where p_L is the wage rate. Note that we only include the items that the household consumes in this constraint, which excludes the “commercial crop,” Q_c (that is, $X_c = 0$).⁴⁴ The farm household’s full-income constraint is the value of its time endowment and any nonwage income,⁴⁵ minus the value of the variable inputs (that is, the inputs whose quantity used can be varied) it uses to produce the farm’s outputs: $Y = p_L T + \sum_{j=1}^M p_j Q_j - \sum_{i=1}^N p_i V_i - p_L L + E$. In this expression, T is total household time, the p_j are the prices the household can get for its outputs,⁴⁶ the Q_j are the amounts of these products it produces, the p_i are the rental prices of variable inputs (equipment, animals, feed, seeds, and so forth), V_i are the quantities of these variable inputs used in production on the farm, L is the total amount of labor used on-farm (family, hired, or both – “hired” labor can be negative if the family is a net supplier of labor to off-farm demanders of labor) and E is the nonhousehold-labor income. Note that the left-hand side of the full-income constraint is the budget constraint and that the full-income constraint is essentially that of the household production model.

We’ll use a slightly different expression for the constraint on production technology than we’ve used so far. The farm’s implicit production function is $G(Q_1, \dots, Q_M, V_1, \dots, V_N, L, K_1, \dots, K_K) = 0$, in which the K_i are fixed inputs such as land and buildings. The implicit production function is a shorthand expression for an array of production activities, since it includes all the outputs Q_j and all the inputs $V_i L$, and K_k .

The farm household maximizes its utility subject to the constraints on its full income (including the budget constraint), the constraints that technology imposes on its production, and to the condition that the prices of all outputs and inputs appear as given to the household. The problem the family is solving is $\mathcal{L} = U(X_a, X_m, X_L) + \lambda[p_L T + (p_c Q_c + p_a Q_a - p_L L - p_v V) + E - p_L X_L - p_a X_a - p_m X_m] + \mu G(Q_c, Q_a,$

$L, V, K)$, which combines the budget constraint with the full-income constraint. The full set of first-order conditions that the family has to manipulate to maximize its utility includes leisure (X_L), the consumption of the home-produced good (X_a), the consumption of the purchased good (X_m), the production level of the home-produced good (Q_a), the production level of the “commercial crop” (Q_c), total labor demand (L), the level of the variable input (V), the implicit (shadow) value of full-income (λ), and the shadow value of production (μ).

From examination of these first-order conditions, we can deepen our insight into what is distinctive about the family enterprise model. The first-order conditions include equating the value of the marginal product of labor to the wage rate: $p_a \Delta Q_a / \Delta L = w$. Now remembering that $\Delta Q_a / \Delta L$ is really $\Delta Q_a(L, V, K) / \Delta L$, this expression for the marginal revenue product of labor can be inverted (solved) as a labor demand equation: $L^d = L(w, p_a, p_v, K)$. The important thing to notice about this labor demand equation is that it is exactly the sort of labor demand equation we would have expected to get much earlier in this chapter, in much simpler models. The only determinants of labor demand are the prices of the product, labor, and the variable input, and the quantity of the fixed factor available; no other choice (endogenous) variables, such as consumption quantities, enter into the determination of the optimal quantity of labor to be used in farm production.

Now turn to the first-order condition for the consumption items: $\Delta U / \Delta X_i = \lambda p_i$ for each consumption good (“ i ” representing, in turn, the purchased good, the home-produced good, and leisure) and $\sum_i p_i X_i = Y^*$, where Y^* is the level of full income associated with the profit-maximizing choice of labor used in production, which is, of course, linked to leisure, X_L . These first-order conditions on the consumption items can be solved to give us demands for each of those goods in the form $X_i^d = X_i(p_a, p_m, p_L, Y^*)$, which should look quite familiar: the demand for any item is a function of its own price, the prices of substitutes and complements, and real income. But remember that Y^* also equals the value of endowments (total time and exogenous income) and farm

profits, which is the value of outputs minus the cost of inputs. Now notice two implications of this fact: first, the production technology embodied in the marginal revenue products of variable inputs, quantities of fixed factors of production, and prices of inputs affect consumption choices via their influence on Y^* ; and second, in the case of the good that is sold off-farm as well as consumed, the usual income effect is now a profit effect with the opposite sign of the usual income effect, dampened by the quantity of the good consumed on-farm.

From the first implication, we get the result that, even with all markets in existence, although consumption decisions do not affect production decisions, production decisions and conditions do affect consumption choices. From the second implication flow a number of quantitative effects on responses to price changes. Negative own-demand elasticities (for example, the elasticity of the quantity of the purchased good consumed, X_a , with respect to its own price, p_m) are smaller negative than in the ordinary consumer case; they can even be positive. Cross-price elasticities (for example, the response in the consumption of X_a as the price p_m changes) that are ordinarily negative (that is, the goods in question are complements) or small negative (modest complements) get smaller negative or turn positive, and positive ones get larger positive. The derived elasticity of labor supply (the response of family labor supply to a change in the output price p_a) turns from small positive to negative, which has a spillover into the labor market. Labor supply elasticities – the response of family labor supply to the wage rate) – can switch from negative (from a backward-bending labor supply curve) to positive or become larger positive.

10.8.2 *Effects of missing markets on labor allocation*

If some market does not exist, the farm household is constrained to equate its consumption of the commodity with its own production of it. In terms of the constrained maximization problem in which we formulated the farm household's utility maximization problem when all markets existed, the household now faces an additional constraint, and the utility maximization itself determines the implicit, or shadow, price of the

good without a market. These shadow prices will be functions of market prices, the household's time endowment and fixed inputs, and either attainable utility or exogenous income. Changes in market prices will now have two routes of influence on allocations: a direct effect and indirect effects via the shadow prices.

We can find out what one of these shadow prices would be by using the expenditure function, which identifies the least expenditure that will yield a given level of utility to a consumer faced with given prices – only one of our prices will remain to be determined. First, we will do the following: we'll set the expenditure function equal to the value of full income of the farm household. In doing this, we'll express the part of the full income that is farm profits as a profit function. Second, we'll take advantage of some useful properties of the expenditure and profit functions: a variation in the expenditure function resulting from a change in one of the prices yields the demand function for the commodity whose price varied, and a similar variation in a price in the profit function yields the negative of the demand function for the input whose price was varied. Third, we can solve the result of the second step for the shadow price of whichever good or factor whose price we've varied. Finally, through some further manipulations on the first step, we can obtain a general expression for how one of these shadow prices changes in response to market prices and family characteristics and assets.

Start with the expenditure function: $e(p_a, p_m, \bar{p}_L, \bar{U}) = \bar{p}_L^* T + \pi(p_c, p_a, p_v, \bar{p}_L^*, K) + E$. The left-hand side of this expression defines an expenditure level just like $\sum_i p_i X_i$ but offers the added advantage of showing just how much of each X_i must be purchased at the set of prices p_i to keep utility constant at the level $\bar{U}$. Over on the right-hand side, the expression $\pi(\blacksquare)$ is the profit function and is just another way of expressing what we did with the notation $\sum_j p_j Q_j - \sum_i p_i V_i - wL$. Now, for step two, we show the expression for how the magnitude of this first expression would change if we changed the wage rate by a small amount: $\Delta e(\blacksquare) / \Delta \bar{p}_L^* = T + \Delta \pi(\blacksquare) / \Delta \bar{p}_L^*$. To explain this expression a bit, on the right-hand side, if we were to vary the wage rate a bit in the first expression, the value of the product $\bar{p}_L^* T$ would just change by the amount of time, T ; similarly,

exogenous expenditures wouldn't change at all because they're not affected by the wage rate, so that term contributes nothing to the results of the wage-rate alteration. Now, $\Delta e(\blacksquare)/\Delta \bar{p}_L^*$ is the demand for leisure (leisure is in the demand function and hence in the expenditure function), and $\Delta \pi(\blacksquare)/\Delta \bar{p}_L^*$ is the marginal product of labor. Think about the latter like this: if we express profits as the difference between the value of output and the cost of inputs and we increase the wage rate a bit, to keep profits constant without changing anything else, we'd have to reduce the labor hours we used.⁴⁷ The relationship between labor used and the wage rate is the demand for labor.

For our third step, we can invert the last expression to solve for the shadow wage rate as a function of the exogenous variables on both sides: $\bar{p}_L^* = \bar{p}_L^*(p_a, p_m, p_c, p_v, K, U)$. This expression tells us which variables influence the shadow wage rate when there is no labor market, but we need to do some further manipulations to understand precisely what their effects are. Go back to the second expression we developed by varying the shadow wage rate – essentially the demand for leisure and the demand for labor. We vary each of the variables in these two functions, one at a time, and with some further manipulations we get $\Delta \bar{p}_L^*/\Delta Z = -(e_{LZ} - \pi_{LZ})/(e_{LL} - \pi_{LL})$, in which the notation e_{LZ} , and so forth, represents $\Delta[\Delta e(\blacksquare)/\Delta \bar{p}_L^*]/\Delta Z$ and Z stands for each of the variables p_a, p_m, p_c, p_v , and K , in turn. Thus, e_{LM} would stand for how the demand for leisure changes in response to a small change in the price of the purchased good, p_m ; similarly, π_{LV} represents how the demand for labor changes in response to a change in the price of the variable input, p_v . The denominator of this expression is known to be negative because of the shapes of the expenditure and profit functions.⁴⁸ The numerator can be of either sign, depending on the relationships between the leisure and labor demand functions and the variable being adjusted. Frequently, we can determine the sign of the numerator if we have some assurance that a pair of commodities or factors are substitutes or complements. Consider a couple of examples. If Z is p_m , the price of the purchased commodity, the numerator is $-e_{LM}$ (p_m does not enter the profit function, so $\pi_{LM} = 0$). If leisure and the purchased commodity are substitutes, $e_{LM} > 0$,⁴⁹

making $-e_{LM} < 0$, altogether creating a positive response of the shadow wage rate to an increase in p_m . Alternatively, if Z is the good both produced and consumed, whose price is p_a , the numerator is $-(e_{LA} - \pi_{LA})$. The first term is positive if leisure and X_a are substitutes (for the same reasoning behind the sign of e_{LM}), and π_{LA} is the response of the output of Q_a to the wage, which should be negative. So the entire numerator is negative, divided by a negative denominator, giving a positive response of the shadow wage rate to an increase in p_a . For yet a third example, this time of an input price, suppose Z represents p_v , the variable input. This price enters the profit function, as π_{LV} , but not the expenditure function. The sign of π_{LV} depends on whether labor and the variable input are gross substitutes or complements. If they're gross substitutes – that is, including the scale effect as well as the substitution effect – an increase in the relative price of the variable input decreases the quantity of the variable input used; alternatively, an increase in the wage rate would decrease the employment of labor – $\pi_{LV} > 0$ and $\Delta \bar{p}_L^*/\Delta p_v < 0$.

The absence of a labor market affects the responses of consumption to own-price changes in ways that differ from their effects when a labor market does exist. The absence of a labor market can convert pure income effects into combinations of substitution effects, because of the change in the shadow price of labor, and income effects. The income effect of a change in the price of the “cash” crop on the demand for leisure is smaller in the absence of a labor market: it raises the shadow wage and induces a substitution away from leisure. Output responses are more likely to be positive the larger the number of variable inputs there are and the greater the substitutability between labor and those other inputs, but they will be smaller in the absence of a labor market than they would be with one because of the rise in the shadow wage. Correspondingly, the response of marketed surplus, $Q_a - X_a$, to its own price, p_a could be positive or negative, with the likelihood of a negative response boosted if X_a and leisure are substitutes, since the shadow wage will rise. Altogether, the absence of a labor market complicates the understanding, and the predictability, of the responses of many allocations that are relatively straightforward in the presence of markets.

10.8.3 Restrictions on household activities

Agricultural households may be forced to work on large holdings owned and managed by others – or find the alternatives unacceptable: collective farms in recent times, temple and large private estates in the ancient Mediterranean basin. In these roles of unequal authority, they may find themselves prohibited from making allocative decisions that would be in their best interest otherwise. They may be required to devote more resources to some activities than they would find optimal and less to others than would be optimal. In some cases, restrictions of this sort can have the effect of prohibiting the emergence of a market in a particular good or service, such as labor, but short of outright prohibition these quotas can impose nonseparability on their production and consumption decisions. In this section, adapted from Sicular (1986), we show the economic structure of such quotas on the farm household and work through the consequences of a few examples.

Let's start from the beginning: the family has its utility function $U(X_1, \dots, X_n)$, which it maximizes subject to its production function, $G(Q_1, \dots, Q_n)$ and its budget constraint, $\sum_i p_i X_i \leq \sum_i p_i Q_i + \sum_j p_j A_j$, where the A_j are endowments of various assets, including equipment, land and time. The Q_i in the production function refer to the same commodities as the corresponding X_i in consumption. With no restrictions on their allocative decisions, the first-order conditions for their maximization problem require that the technical rates of substitution in production and the corresponding marginal rates of substitution in consumption equal the ratios of product prices: $(\Delta G / \Delta Q_i) / (\Delta G / \Delta Q_j) = (\Delta U / \Delta X_i) / (\Delta U / \Delta X_j) = p_i / p_j$. These conditions will change with the distorting effects of quotas.

Quotas can be characterized similarly to marketed surplus, as the difference between production and household consumption, $S_i = Q_i - X_i$, with an allowance for the existence of pre-existing stocks.⁵⁰ Additionally, two types of quota can be imposed, one as an absolute limit on household purchases and sales, the other as limitations on the transactions in one good

linked to transactions in some other good. The fixed quota specifies an absolute upper or lower limit, while the latter, the variable quota, may specify upper and lower percentage limits or tie sales or purchases in one market to sales, purchases or production in another. Of course, quotas on net sales can be either positive for sales or negative for purchases.

A fixed quota can be described by $S_i = A_i + Q_i - X_i \geq \bar{S}_i$, which says that the household's net sales of good i , which is its initial endowment of the good, plus its current production of it, minus its own consumption of it, must exceed the quota level $\bar{S}_i$. If $\bar{S}_i > 0$, net sales must exceed the minimum level; if $\bar{S}_i < 0$, net purchases (negative sales) must be smaller in absolute value than the specified level. If $\bar{S}_i = 0$, no trade in good i is permitted. Three types of variable quotas can be imposed. First, the sales quota for good i can be tied to its production: $S_i = A_i + Q_i - X_i \geq \Theta Q_i$, which requires that net sales of good i be at least the fraction Θ of the household's production of it. A value of $\Theta = 1/2$ would mean that the household must sell at least half of its output to the authority imposing the quota. Similarly, the sales quota for good i could be tied to the production of some other good j : $S_i = A_i + Q_i - X_i \geq \Theta Q_j$. Again, purchase quotas would be mandated by a negative Θ . Finally, the household's sales quota for good i could be tied to its sale of some other good: $S_i = A_i + Q_i - X_i \geq \Theta (S_j = A_j + Q_j - X_j)$.

A household facing one of these types of quota has another constraint on its household utility-maximization problem. For example, imposition of a fixed quota would leave a household with the following Lagrangean: $\mathcal{L} = U(X_1, \dots, X_n) + \lambda [\sum_i p_i (X_i - Q_i - A_i)] + \mu G(Q_1, \dots, Q_n) + \varphi (\bar{S}_i + X_i - Q_i - A_i)$. To maximize this utility, the household chooses the production levels of all n goods, Q_i ; its consumption levels, X_i , which would include leisure and therefore labor supply; and the shadow values of income (λ), of profits (μ), and of the quota itself (φ). The ratio φ / λ is the change in household profits following a decrease in the quota level. In the case of this fixed quota, the marginal rate of substitution in consumption and the technical rate of substitution in production are modified by the

additional constraint: $(\Delta G/\Delta Q_i)/(\Delta G/\Delta Q_j) = (\Delta U/\Delta X_i)/(\Delta U/\Delta X_j) = (p_i + \varphi/\lambda)/p_j$, while between other pairs of goods not directly affected by the quota on good i , $(\Delta G/\Delta Q_i)/(\Delta G/\Delta Q_k) = (\Delta U/\Delta X_j)/(\Delta U/\Delta X_k) = p_j/p_k$. In the case of this particular quota, the price distortion facing households as consumers is the same as the distortion facing them as producers, but that will not be the case with the tied quotas. If $\bar{S}_i > 0$ and sets a floor on household sales of good i , a decrease in S_i will increase the household's profits, and $\varphi/\lambda \geq 0$. In such a case in which a household is forced to sell more than it finds optimal (or, in reverse, must buy more than it wants, for $\bar{S}_i < 0$ placing a ceiling on purchases), a binding quota ("binding" meaning that in the quota's absence the household would behave differently) will cause the shadow price of good i to rise above its market or administrative level p_i . Alternatively, if $\bar{S}_i > 0 > 0$ and sets a ceiling on sales, a decrease in the quota will reduce household profits; $\varphi/\lambda < 0$ and the shadow price of good i is lower than its administrative price. Similarly if $\bar{S}_i < 0$ and sets a floor on purchases of good i .

With the first tied quota, in which the sales quota is linked to the household's output of the same good, the consumption shadow price for good i is $\bar{p}_i^c = p_i + \varphi/\lambda$ and all other consumption prices are unaffected, while the production shadow price for good i is $\bar{p}_i^p = p_i + (1 - \Theta)\varphi/\lambda$. The price of the good to which the sales quota is tied is distorted to a lesser extent, governed by the tying ratio Θ . In fact, if $\Theta = 1$, meaning that the entire current output must be sold to the quota authority, the production shadow price falls back to its undistorted level: $\bar{p}_i^c = p_i$. By looking at the quota tied to own production, you can see that when the tying ratio, Θ , equals 1, the production level effectively falls out of the quota, which becomes a restriction that current consumption of good i be no greater than the stocks of good i held at the beginning of the period: $X_i \leq A_i$. The tying ratio requires that all current production be sold to the quota authority (or whoever the quota authority designates), leaving any consumption of the good by the household producing it coming out of stocks they held prior to the imposition of the quota. As stocks dwindle to zero, the value of φ in the ratio φ/λ would tend to infinity,

indicating that the shadow price of the good rises without limit as its availability goes to zero.

When the sales quota on good i is tied to the production level of good j , the shadow consumption price of only good i is affected but the shadow production prices of both goods i and j are distorted, but in opposite directions: $\bar{p}_i^c = p_i + \varphi/\lambda$ while $\bar{p}_j^c = p_j$; in production, $\bar{p}_i^p = p_i + \varphi/\lambda$ and $\bar{p}_j^p = p_j - \Theta\varphi/\lambda$. The quota is adjusted by changing the tying ratio, Θ . In this case, if the quota sets a binding floor on sales, a higher tying ratio depresses the shadow production price of the other good, reducing the incentive of the household to produce it while reinforcing the incentive to produce the quota good. Intuitively, the household can avoid some of the sales it doesn't want to make by reducing its output of the good to which its forced sales are tied as a fixed percentage. If the quota imposes a ceiling on purchases of good i , limited by $\bar{S}_i < 0$, an amount below what the household would like to buy, the ratio of Lagrange multipliers φ/λ is also positive (indicating that raising $\bar{S}_i$ would improve the profits and utility of the household by permitting the greater purchases of something they want to purchase), and the shadow production price of good j falls below its administrative or market price, reflecting the household's reduced incentive to produce it.

Tying a sales quota for one good to the sales of another good affects the shadow consumption and production prices of both goods. Thus, if the sales quota for good i is tied to the household's sales of good j by the proportion λ , the shadow consumption prices for i and j are $\bar{p}_i^c = p_i + \varphi/\lambda$ and $\bar{p}_j^c = p_j - \Theta\varphi/\lambda$, and their shadow production prices are identical. In cases in which a reduction in the quota would raise the household's utility and profit, that is, $\varphi/\lambda > 0$, this quota on good i raises its shadow price above the good's administrative price and depresses the shadow price of the good to whose sales it is tied. The tying proportion, Θ , does not affect the shadow price of good i (the one with the quota imposed on its sales) but depresses the shadow price of the good to whose sales it is tied. If the household has to sell more than it wants of good i because of the amount of good j it is selling, the relative shadow values of i and j will

change to reflect the relative scarcity imposed on i and the relative abundance imposed on j by the tied quota.

In thinking about the consequences of these quotas it is useful to recall that the Q_i and X_i in terms of which these quotas have been defined are themselves functions, the Q_i governed by their supply functions and the X_i by their demand functions. The quotas can't be imposed without influencing a lot of allocations that the agency imposing the quota has no interest in influencing, or might even want not to influence. From the first-order conditions that gave us the ratios of marginal conditions in production and consumption, we also can derive profit functions and demand functions, and from the profit functions can be derived supply functions. The supply functions for the Q_i are positive functions of those goods' shadow prices to the household while the demand functions are negative functions of the shadow consumption prices. The demand functions for goods not directly affected by a quota contain the shadow consumption prices of goods that are affected by quotas, and of course, the demand functions for the goods on which quotas are imposed specify the substitutability and complementarity between the quota good and all other goods, including leisure. And in terms of separability of production and consumption decisions, quotas definitely impose nonseparability: production decisions become influenced by utility considerations as well as vice versa.

10.8.4 *Implications of the family farm model*

This entire section has demonstrated the pervasiveness of logically accountable actions and forces producing counterintuitive allocative consequences at the level of the individual family. Does this mean that the use of the simpler, more "standard" models of supply and demand for entire sectors and economies will yield predictions in such striking contrast to actual allocations as to be worthless? In general, the answer is "no." Such an inference would involve the fallacy of composition. Different families or other production units, faced with external ("market") prices would find different levels of relative prices that would tip them from, say, negative

to positive price responses. Other agents would find it worthwhile to enter production of some good when its price reached a high enough level, regardless whether individual family producers responded similarly at that level of price. Having issued this general warning, it is nevertheless important to recognize when a set of constraints applies to an entire class of production, such that all producers, actual and potential, face similar incentives relative to their incentives in other activities. Exchange rate controls and tariffs, particularly on intermediate goods – those used in the production of final goods – are well known for their capacity to induce allocative responses that are not straightforward.

Having illustrated the complications caused by the absence of labor markets and distortions in other markets, it is useful to discuss the empirical incidence of such characteristics, at least in recent and contemporary developing countries, which may offer our closest correspondences to technological and contractual conditions in the Mediterranean basin in antiquity. While the notion that rural labor markets generally were absent from those economies was widely held from the 1940s well into the 1960s by scholars who hadn't checked carefully, closer investigation of rural areas, beginning in the 1970s (and retrospective investigations as well) consistently indicated thriving labor markets, with people engaging their labor services under a variety of contract types even within the same village. There also has been little evidence found of important departures from competitive conditions in those labor markets – that is, little evidence of collusive behavior by employers or monopsony. These findings correspond well with theoretical information developed during the same period that the "large numbers" of buyers and sellers required for perfect competition are surprisingly small. Thus, it doesn't take all that many employers in a rural area to keep one another pretty competitive (Binswanger and Rosenzweig 1984, 3, 23, 35).

This said about the extensiveness of competition in contemporary rural labor markets, it also is true that when some markets fail to exist, which can happen particularly easily for activities in which outputs follow the application of inputs with an appreciable time lag, transactions for one class of goods or services can be piggybacked with transactions in other goods or services.

These transactions that require the spanning of time generally involve credit and insurance. Thus, it is not all that rare to find credit and *de facto* insurance extended to individuals by their employers since the employment relationship provides a means of revealing information about borrowers as well as instruments for controlling aspects of their behavior that would affect their repayment probabilities.⁵¹

10.9 Slavery

Slavery is not a topic about which economics has much in the way of theory. Slavery is a variegated array of institutional structures that have appeared frequently over time and around the globe. The principal common theme across the different instances of slavery is that the slaves generally have greater restrictions on many choices than do other individuals nearby, who may be called “free” in contrast, although the meaning of “freedom” also has been subject to institutional negotiation. Although there is no commonly accepted “economics of slavery,” slavery is a natural topic to which to apply models from capital theory and labor economics. I take that approach to this important subject of antiquity, offering something like a checklist of considerations that a scholar of Mediterranean antiquity may want to remember when thinking about slavery. The names of the subsections should come as no surprise to any reader who has persevered this far.

Economic aspects of ancient slavery, particularly Greek and Roman, Mesopotamian and Egyptian somewhat less so, have been studied extensively, if with limited success. To demonstrate examples of successes and limitations, consider two prominent recent papers by Scheidel (2008, 2012). First, both papers give extensive attention to slaves’ incentives – to work, to shirk, to run away – an important topic when considering the use of slave labor. To focus the assessment of incentives, Scheidel employs a typology of work and incentive systems (effort-intensive work relying on pain incentives and care-intensive work relying on ordinary rewards) from a paper by Fenoaltea (1984) and finds instances where both pairings are found in the Graeco-Roman world, which, while it may provide a new viewpoint,

primarily offers new names to well-known phenomena. However, while relying on the Fenoaltea typology in both papers, the earlier paper offers a recitation of its inaccurate predictions (Scheidel 2008, 108–109).

Second is an alternation between presentation and analysis of price data and characterization of some situations as labor shortages which the operation of a price system would eliminate, if not necessarily to all parties’ liking. To begin with the prices, Scheidel (2008, 124, Table 4.4; expanded somewhat from Scheidel 2005a, 10, Table 3) compares estimates of slave prices and wages of nonslave unskilled labor around the Graeco-Roman Mediterranean, using regional price and wage levels, both expressed in tons of wheat equivalents, as indicators of regional scarcity of labor. A peculiarity of those figures that goes unremarked is the relationship of an unskilled rural worker’s annual wage payment to the full price (wage payments capitalized over the expected working life) of the slave. For Classical (5th – 4th century B.C.E) Athens a mean slave wheat price of 1.4 tons and a mean annual wage of an unskilled rural worker of 0.4 ton would imply a 25% discount (interest) rate if the slave were expected to live (be productive) for 5 years, 13% for 10 years, and 8% for 20 years;⁵² not an unreasonable range of discount rates. However, in the case of third-century C.E. Egypt, with average slave price of 4 and annual unskilled rural wage of 3, a buyer with a 25% discount rate would have expected a slave to survive for a year and about 3 weeks; at a 10% discount rate, about 2 years and 10 months; and with the circumstances of the Roman price edict of 301 C.E., slave prices of 2.75 and annual wage of 2.25, a buyer with a 10% discount rate would expect a slave to last just 2 years. These implied expectations seem considerably too low. Nonetheless the assembly of price data is an excellent initial step in economic analysis. However, in the same papers, Scheidel also follows the widely employed habit of referring to the supply of free labor as “insufficient or otherwise inadequate” (2012, 96) or “labor shortages” (2008, 118). In the Roman case, the lengthy mobilization of the Roman Army from the Republic into the Early Imperial period, à la Hopkins (1978), is the widely accepted cause (demand shift) of the “shortage” of civilian labor providing an (at least partial) explanatory device

for the existence of slavery in various societies, and Scheidel appears to accept this explanation and even extend it to various Greek cases.⁵³ Scheidel builds an interesting case for a military demand shift in the Greek case, but does not address why prices (wages) did not adjust more easily when chattel slavery was expanded from its relatively domestic, Homeric version to the economy-wide institution it became by the time we have some written evidence. Linking these arguments more explicitly to relative prices could produce a better understanding of ancient social phenomena and choices. For a related example, while the Laureion mine slaves are a popular example of Greek slavery, to designate slaves as “essential” in those mining operations (Scheidel 2008, 106), which may seem plausible without deeper consideration, implicitly designates a particular form of production function, possibly fixed-coefficients. But those mines had been worked since the Bronze Age; do we really believe that slaves were employed in those mines at that time? If not, or if we’re not sure, then their essentiality in the fifth century B.C.E. and later surely should come under question.

Third are the appeals to tastes and habitual behavior (Scheidel 2008, 125) and institutional acceptability – tastes by another name – as a “precondition” for extensive adoption of slavery in a society (Scheidel 2011, 96). Either version of the taste explanation for economic behavior can produce either an infinite regress as an explanation (it had been done that way for a long time) or no explanation at all.⁵⁴

Fourth is a combination of taxonomizing and a search for origins. Giving a name to a type of economy – slave economy in this case (Scheidel 2008, 105; 2012, 89) – runs somewhat cross-grained to economic analysis, the operation of which is founded on the concept of identical forces operating through a variety of institutional settings that may affect outcomes but not the basic forces themselves. This said, Turley’s (2000, 4–5) tripartite checklist of societal characteristics for distinguishing societies with slaves from slave societies (interchange “economies” with “societies”), which Scheidel adopts, offers a more comprehensive coverage of individuals’ involvement with slavery than Hopkins’ (1978, 98–102) admittedly arbitrary cut-off point of slaves’ comprising 20% of a population.

Rather than exploit Turley’s incentive-based checklist, Scheidel (2008, 115–116) develops his own “preconditions” for the “emergence of large-scale slavery” – essentially demand and supply, subsequently expanded to include institutional acceptability, or tastes (Scheidel 2012, 96). The application of this origins checklist to early Mesopotamia and Old Kingdom Egypt assumes that “lack of labour shortages” in both cases kept slavery to minimal proportions, and that when slaves became more extensive in Mesopotamia during the Ur III period, demand apparently increased (Scheidel 2008, 120–121). Circularity aside, such a supply-demand checklist could yield false positives, which could be corrected each time by bringing in tastes. Nonetheless, comparative analysis of Greco-Roman slavery with the situations in Mesopotamia (early and late) and Pharaonic and later Egypt could prove interesting. Prisoners and other captives from wars were a common source of slaves throughout Mesopotamian antiquity (Siegel 1947, 9–11; Turley 2000, 35), and at least in Neo-Assyrian times, slaves in Mesopotamia were sprinkled throughout the occupations of the economy as broadly as were free workers (Warburton 2005, 198; Dandamaev 2009), while in Egypt, although entire populations of captured adversaries could be enslaved immediately after capture, widespread slavery never caught on (Bakir 1952/1978; Eyre 2010, 302; Kehoe 2010, 321–322). While searches for origins of economic practices are often elusive, sharper analytical comparisons among these cases should be productive regarding causes, operations, and consequences.

Scheidel’s contributions to the study of slavery in the ancient Mediterranean have been substantial, interesting, and provocative. He has assembled a wide array of scattered information not necessarily related heretofore and has employed some basic economic reasoning to tease some new understandings from it. In the endeavor he has sometimes raised as many questions as he has provided lasting answers, but such is the nature of the business.

10.9.1 *The supply of slaves*

Remember that a supply curve (or function or schedule, or in situations in which those terms are technically incorrect, just a “relationship”)

relates the quantity supplied to the price offered: the higher the price, *ceteris paribus*, the greater the supply. Xenophon appears to have recognized this relationship (Westermann 1955, 9). The supply of slaves is the supply of a stock, not of a flow such as the supply of labor by slaves or free workers would be. As such, during periods of lower interest rates, people could pay more for a slave, and vice versa during periods of higher interest rates – *ceteris paribus* of course.

Naturally, supply curves for slaves in various parts of the ancient Mediterranean, at various times, would be difficult if not impossible to estimate. Conversion of prisoners of war into slaves is admittedly difficult, although probably not impossible, to cram into a supply framework. However, the concept of supply remains useful for thinking about the scale of acquisition of slaves: for example, if the price of slaves at a central market, say Delos, were higher, Cilician pirates could have afforded to reach deeper inland in raids for captives. Such a supply curve would not distinguish individual characteristics of slaves (and whether civilian slave capturers such as pirates looked for individuals with particular characteristics or just bagged an entire lot and threw the apparently most worthless overboard to avoid feeding them may also be unknown) but would simply identify aggregate numbers at different prices. The literature on ancient slave supply tends to address geographical origins of slaves, their method of acquisition (capture, natural reproduction, debt bondage, and so forth), and individual prices observed for particular individuals (or averages of individuals) at particular places and times in antiquity. This price information may tell us something about the labor market equilibrium at the time, or it may be contaminated by characteristics of individual slaves auctioned or otherwise sold, such as strength, agility, skills, looks, but it does not contain information on the responsiveness of the number of slaves offered at different prices. Deliberate slave breeding may well have been responsive to prices, or may not have, since the time between the delivery of a slave with characteristics commanding current prices could be 15 to 20 years. One might think that wars would not be responsive to slave prices, but Braund (2011, 118) suggests that the Spartan invasion of Asia Minor in the early 390s B.C.E. anticipated “the prospect of the enslavement of

their barbarian enemies,” although it is not clear that the Spartans thought closely about quantity versus effort.

Let’s nonetheless turn to the standard topics treated under slave supply in the literature. “Where did the slaves come from?” is a natural first question, and we must think about this issue differently from the issue of traditional labor supply, which will be a topic in slaves’ incentives in subsection 10.9.5. The two principal methods of supply of slaves are domestic production and imports. Domestic production of slaves amounts to either deliberate breeding of slaves or a somewhat more incidental breeding fostered by the establishment of slave families.⁵⁵ Breeding, or even permitting slaves to have families that would maintain any given slave owner’s stock of slaves, likely would depend on the relative costs of breeding versus importing. Raising a slave from infancy to a time when he or she can begin to produce more than what is required for subsistence may be substantial when compared to the discounted value of the adult earnings. If slave productivity were sufficiently low, a slave-using producer would want to keep only adults, preventing their reproduction through whatever means were ethically acceptable at the time. If this sounds like it would make slaves into automatons, and their owners into unthinking assembly-line operators, we will return to several themes surrounding this subject when we discuss slaves’ incentives. The point remains, however, that the relative cost of producing slaves domestically versus importing them through the various means discussed below, would have been an important influence on the mix of supply categories.⁵⁶

Imports, as slaves, of people who were not born slaves, require the transition from nonslave to slave at some point, either immediately prior to sale as a slave or later in that person’s post-sale life. The “transition,” as we rather clinically describe it, would likely involve violence – capture in war or just plain capture – although we cannot rule out the institutional possibility that people at various times and places have been able to sell themselves or members of their family into slavery to settle debts, or in the case of other family members, possibly just to have a bit of extra consumption. Imports require exports, even in the case of capturing prisoners of war – just think of all the resources poured into the equipping

and feeding of the Roman infantry and cavalry (service exports) who sent back slaves from northern and eastern Europe. Of course, the same accounting applies to the import of prisoners of war in the ancient Near Eastern empires. Even the New Kingdom Egyptians' capture of Libyans and various Sea Peoples within the territory of Egypt is subject to this type of accounting because the Egyptians surely expended considerable resources to repel the invaders, some of whom were captured.

The slaves captured not in war but by what would have been called slavers in the seventeenth and eighteenth centuries C.E. in the Atlantic probably were harvested by groups contemporarily called pirates – and even called pirates by the ancients at various times, according to political expediency (Braund 2011, 120–121; Scheidel 2011, 297–298). We could call these “unofficial imports,” in contrast to the “official imports” of prisoners of war. Such terminology distinguishes between the industries or sectors involved in the “production,” or “harvest,” or “supply,” and very likely may give hints to differences in resource costs of supply.

Another aspect of the supply of slaves that bears attention is their marketing: moving the supplies from initial locations to final demand sites. Delos in the second and early first centuries B.C.E. performed such a function. Prices gave signals that indicated where the supplies should be distributed. The apparent production points of slaves (or of people who became slaves) were widely separated from the locations where the slaves were used, and the slave-marketing industry, including transportation and information provision, must have been large.

10.9.2 *The demand for slaves*

The obvious alternative to slave labor was free labor, whatever that contemporary term may have meant at various times and places. In fact, the slave versus free dichotomy may be too simple for parts of ancient Mesopotamia, and even certain periods in Greece and Italy. Rights to various actions – or to the decision to take or not take certain categories of actions – may have varied more on a continuum than the relatively stark conditions of slavery in the southern United States from the seventeenth to the nineteenth

centuries C.E. have encouraged scholars to contemplate. Whatever those gradations of rights between certainly free and certainly unfree, it is useful for the contemporary scholar to think in terms of the substitutability between labor with one set of rights and labor with other sets of rights. Doing so presents us with the potentially less intractable question, “Why use group *X* (“slave”) rather than group *Y* (“free”)?”⁵⁷

It is easy to recognize that a rightward shift in the demand curve for labor would drive up the cost of using labor. Attracting slave labor might have been cheaper than attracting freely migrating labor in sufficient quantities to move the labor supply curve to the left *pari passu* with the demand curve. Nonetheless, this framework shifts the focus to why the labor demand curve shifted in the first place? Might new land have been opened? Why, if there was not pressure of a growing labor supply in the first place? Possibly an increase in the demand for something that could be grown on newly opened land, some product that might be exported? If consumed domestically, where would the extra production have come from to pay the additional resource costs of the slave labor used in production? Demands from the military as in the case of Republican and Early Imperial Rome?

Alternatively, might some work have been sufficiently “dirty” that it cost “too much” to pay free labor to do it? The Laureion silver mines in Attica come to mind, with their dreadfully used slave-labor force. Even in such a case, the demand for the product must be sufficient to pay for the additional labor, even if the labor is paid only its bare subsistence cost, all the producer’s surplus in the “wage” being siphoned off by the slave owners. If slaves did not at least pay for themselves in the sense of providing their owners a return at least equal to their food, shelter, clothing, and any other expenses they might have incurred – medical attention, transportation costs, and so forth – they would have been consumer goods, comparable to fine cloth, imported ivory mirror handles, and so forth, not producer goods.

The Marshallian laws of derived demand provide a useful set of criteria to confront proposed ideas about the use of slaves. Those four important parameters were the demand for the product,

the substitutability for other inputs, the supply elasticity of other inputs (for example, equipment or free labor), and the cost share in production. The demand elasticity parameter tells us to focus on the characteristics of the production activities in which slaves were used. Substitutability for other inputs suggests we look at the relative costs of other inputs that might have been used in coordination with slaves and the technologies embodied in those other inputs and their methods of combination. The supply elasticity leads us to ask about the relative availability of those other factors of production. The cost-share parameter suggests that we look at how important slave labor was as a cost share in whatever production activity it was being used in. A low cost share indicates some ability for the slave owner to squeeze some extra rents out of the final product for the use of his slaves, with the slaves possibly being able to bargain for some of them. A high cost share suggests that keeping the cost of the slave labor contained would have been important, with less happy consequences for the slaves involved.

10.9.3 *Investment in slaves*

Just buying a slave is an investment from the perspective of the purchaser, and the slave has all the economic characteristics of a piece of equipment: an expected productive lifetime (including the likelihood of escape), an anticipated marginal revenue product at each point in his or her lifetime, and a rate of return that would make the purchase just profitable. The life expectancy of a slave would have had a distribution, just as a lightbulb has a distribution of expected hours of service, and the distribution of slave lifetimes may have been no harder to predict than the economic lifetimes of other factors of production – livestock, wooden or metal equipment, ships, wagons, and so forth. Consequently it may be unreasonable to say that it would have been riskier to invest in the purchase of a slave than in the purchase of a wagon or a share in a ship. Loans to purchase slaves may have been no riskier than loans to purchase other types of capital; it's simply an empirical question.⁵⁸

The other type of investment in a slave is investment in human capital improvements – training. In terms of taking care of his investment, a slave owner might have been able to ensure the type of

care a slave received more readily than he could a free worker. The investment in training a slave would also not be lost to a job change without the employer's permission. To the extent that slave owners had better access to capital than did the average free worker, they could have financed training more efficiently than could free workers, possibly leaving the distribution of skills higher among slaves than among free workers.

10.9.4 *Market consequences of slaves*

The primary market consequence of slaves' contribution to the aggregate labor supply curve is what they do to the wage for all labor – free and slave. At the margin, of course, the free and slave wages should be equal, although the owners of slaves may be able to extract some portion of any producer's surplus in the wages generated by slaves. This latter effect is a worthwhile topic in its own right, but at first approximation the answer to it is limited by the supply curve of slave labor. A relatively elastic (that is, pretty flat) supply curve of slave labor (different from the supply curve of slaves) would generate little producer surplus to be extracted.

Another aggregate consequence of the use of slave labor could be its influence on aggregate savings and investment. Depression of labor income would reduce savings by itself, but extraction of labor income from slaves to slave owners could either increase or decrease saving, depending on the spending patterns of the slave owners. Luxury consumption habits would divert funds away from savings, but thrifty ones could enhance savings by more than the depression of labor income dampened it. A large increase in the labor-capital ratio (slaves are still labor, even if they're capital assets of their owners – we're all our own capital assets today) would raise the return to investment in equipment, and possibly land as well, making investment and hence saving more attractive. These are all empirical questions.

10.9.5 *Slaves' incentives*

By slaves' incentives, I refer to their motivations to provide specific characteristics of work, such as effort, care, attention, diligence, attendance – the characteristics that turn any given number of hours into different quantities of output. It has

been easy to discuss slaves and slave owners as if they were devoid of the usual human characteristics of personality and motivations. Recent research in North American Colonial and United States slavery has pried open the issue of personal interaction between slave labor forces and slave owners, providing evidence of at least implicit bargaining regarding working conditions, even if the slaves' scope for bargaining had obvious strictures. While it is difficult to discover ancient slaves' – or even many recent ones' – individuality and personalities in the remaining forms of evidence, the economic approach to the study of labor points to the different ways that people derive payment for their work, the pervasiveness of their incentives and ability to shirk, and the limited ability of monitoring to elicit good-faith effort. Surely slavery has been replete with efforts to get large numbers – or even small numbers – of slaves to cooperate in order to make harsh conditions less harsh in return for making a product less small. We note some of the scope for such washing of one hand by the other within the limitations of slave conditions.

Generally, the ancient records indicate the considerable dependence of slaves' incentives on the details of legal institutions.⁵⁹ In some times and regions, slaves were able to purchase their freedom – surely an incentive that slave owners could put to their own use as well as the slave's. Slaves' rights to ownership of property, and the security thereof, would have been another such incentive. The legal protection afforded slaves' persons and lives, as well as any property they may have owned, certainly would have comprised part of their incentive structure. Possibilities for emancipation, or at least retirement – such as Eumaeus received (a pension with light work) in the *Odyssey* – would have formed some incentives that owners could have manipulated to the improved satisfaction of both parties. Of course, we might say that the owner could have reneged on promises, but owners would face reputation problems with their slaves as well as do contemporary employers: slaves would be unmotivated by promises that have lost their credibility.

Opportunities for marriage and procreation, subject to the returns in that for the owner noted

above, may have provided incentives, possibly dampened by the potential loss of spouse and children to sale. The variable status of children of one free parent and one slave parent surely affected one or both spouses' incentives for both marriage and procreation, although the possibility of losing such offspring to freedom may have affected owners' willingness to permit certain categories of slave-free marriages.

Finally, returning to the theme of negotiation with which we opened this subsection, negotiating capacity is constrained – or, conversely, conferred – by considerations of supply, demand, and substitutability. The concept of the threat point is useful in analyzing negotiation: if we can't reach an agreement, what can each of us do? Divorce if we're spouses? Walk away if we're business partners? The slave can't walk away, but if he's quick, he can run away, which is a clear threat to slave owners. Even if runaway slaves can be caught, the owner may face a high cost of making an example of them. They have an investment in them, and killing them or maiming them sufficiently to impair their productivity is destructive of the owner's interest as well. Of course, the cost of replacement – or the cost of nonreplacement – affects the comparable threat point on the owner's side. Collectively, slaves hold the threat of revolt, as they demonstrated on notable occasions throughout antiquity (Westermann 1955, 41–42, 64–65, 75).

10.10 Suggestions for Using the Material of this Chapter

Labor being largely ephemeral in the archaeological record, ancient historians and philologists may be able to make more direct use of the models of this chapter. That said, there are still lessons for archaeologists as well regarding how people decide how much to work.

Work, play, and activities between those two poles and on the far sides of both all involve labor, so the scope for applications of this chapter's material to the resource-allocating aspects of people's decisions about their activities should be both wide and deep. Many applications of this chapter's material by ancient historians and

philologists in particular, but archaeologists as well, could derive from the household production concept, which offers methods for associating people's decisions in apparently disparate realms, say between ritual activities and production activities or between household activities and market activities. Similarly with the family enterprise model, as much of the production activity in antiquity was conducted at the level of the family. Both models offer analytical frameworks for understanding how decisions in one realm affect or imply decisions in other realms.

All labor embodies some human capital – what people know about making something in particular and about stuff in general. Farmers, by far the predominant component of ancient labor forces, had much location-specific human capital; they stood to lose much of this wealth if they moved very far. When identifying suspected invading groups into a region, scholars should keep in mind the ability of a new group to get the most out of a new environment. Displaced groups can be expected to have had similar adaptation problems. Newly migrated populations can be expected to display greater productivity

within several generations but the improvements would be expected to attenuate eventually. Some technological changes developed by new populations could reflect local applications of region-specific human capital different from that of the pre-existing population's.

Personal migration is occasionally a phenomenon in antiquity that begs for explanation. In using migration models from development economics, watch for wage effects. Beware of models that assume a fixed wage in a city and let the unemployment rate equilibrate migration.

The theory of compensating differentials suggests that we should expect pay differentials for particularly dirty or uncomfortable or dangerous occupations. The risk of dying in some occupations should be compensated to some extent with a higher wage, with the difference between a safe occupation and the risky one being related by the probability of various types of accident occurring in the risky one. In societies using slave labor, we should expect such occupations to be filled disproportionately with slaves, with slave owners picking up the rents of compensating differentials in wages they don't have to pay slaves.

References

- Archibald, Zosia H. 2011. "Mobility and Innovation in Hellenistic Economies: The Causes and Consequences of Human Traffic." In *The Economies of Hellenistic Societies, Third to First Centuries BC*, edited by Zosia H. Archibald, John K. Davies, and Vincent Gabrielsen. Oxford: Oxford University Press, pp. 42–65.
- Ault, Bradley A. 2005. *The Excavations at Ancient Halieis. Volume 2. The Houses; The Organization and Use of Domestic Space*. Bloomington IN: Indiana University Press.
- Bakir, Abd el-Mohsen. 1952/1978. *Slavery in Pharaonic Egypt*. Cairo: Service des Antiquités de l'Égypte.
- Barnum, Howard N., and Lyn Squire. 1979. *A Model of an Agricultural Household: Theory and Evidence*. Baltimore MD: Johns Hopkins University Press.
- Becker, Gary S. 1974. "A Theory of Social Interactions." *Journal of Political Economy* 86: 1063–1093.
- Becker, Gary S. 1991. *A Treatise on the Family*, 2nd edn. Cambridge MA: Harvard University Press.
- Behrman, Jere R., 1988a. "Nutrition, Health, Birth Order and Seasonality: Intra-household Allocation in Rural India." *Journal of Development Economics* 28: 43–63.
- Behrman, Jere R. 1988b. "Intra-household Allocation of Nutrients in Rural India: Are Boys Favored? Do Parents Exhibit Inequality Aversion?" *Oxford Economic Papers* 40: 32–54.
- Behrman, Jere R. 1997. "Intra-household Distribution and the Family," In *Handbook of Population and Family Economics*, Vol. 1A, edited by Mark R. Rosenzweig and Oded Stark. Amsterdam: North-Holland, pp. 125–187.
- Behrman, Jere R., Robert A. Pollack, and Paul Taubman. 1982. "Parental Preferences and Provision for Progeny." *Journal of Political Economy* 90: 52–73.
- Bell, Clive. 1988. "Credit Markets and Interlinked Transactions." In *Handbook of Development Economics*, Vol. 1, edited by Hollis B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland, pp. 763–830.
- Bergstrom, Theodore C. 1996. "Economics in a Family Way." *Journal of Economic Literature* 34: 1903–1934.

- Bergstrom, Theodore C. 1997. "A Survey of Theories of the Family." In *Handbook of Population and Family Economics*, Vol. 1A, edited by Mark R. Rosenzweig and Oded Stark. Amsterdam: North-Holland, pp. 21–79.
- Biddle, Jeff E., and Daniel S. Hamermesh. 1990. "Sleep and the Allocation of Time." *Journal of Political Economy* 98: 922–943.
- Binswanger, Hans P., and Mark R. Rosenzweig. 1984. "Contractual Arrangements, Employment, and Wages in Rural Labor Markets: A Critical Review." In *Contractual Arrangements, Employment, and Wages in Rural Labor Markets in Asia*, edited by Hans P. Binswanger and Mark R. Rosenzweig, New Haven CT: Yale University Press, pp. 1–40.
- Birdsall, Nancy. 1988. "Economic Approaches to Population Growth." In *Handbook of Development Economics*, Vol. 1, edited by Hollis B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland, pp. 477–542.
- Bondi, Sandro Filippo. 1999. "The Course of History." In *The Phoenicians*, edited by Sabatino Moscati. New York: Rizzoli, pp. 30–46.
- Borjas, George J. 1998. "To Ghetto or Not to Ghetto: Ethnicity and Residential Segregation." *Journal of Urban Economics* 44: 228–253.
- Borjas, George J. 2000. "Ethnic Enclaves and Assimilation." *Swedish Economic Policy Review* 7: 89–122.
- Braund, David. 2011. "The Slave Supply in Classical Greece." In *The Cambridge World History of Slavery*. Volume I. *The Ancient Mediterranean World*, edited by Keith Bradley and Paul Cartledge. Cambridge: Cambridge University Press, pp. 112–133.
- Brunt, P.A. 1980. "Free Labour and Public Works at Rome." *Journal of Roman Studies* 70: 81–100.
- Burford, Alison. 1993. *Land and Labor in the Greek World*. Baltimore MD: Johns Hopkins University Press.
- Cahill, Nicholas. 2002. *Household and City Organization at Olynthos*. New Haven CT: Yale University Press.
- Cahill, Nicholas. 2005. "Household Industry in Greece and Anatolia." In *Ancient Greek Houses and Households*, edited by Bradley A. Ault and Lisa C. Nevett. Philadelphia PA: University of Pennsylvania Press, pp. 54–66.
- Caruso, Ada. 2013. *Akademia: archeologia di una scuola filosofica ad Atene da Platone a Proclo (387 a.C. – 485 d.C.)*. SATAA: Studi di Archeologia e di Topografia di Atene e dell'Attica, 6. Paestum: Edizioni Pandemos.
- Chayanov, A.V. 1925. *Organizatsiya krest'yanskogo khozyaistva [Peasant Farm Organization.]* Moscow: Cooperative Publishing House. [In Russian.]
- Cohen, Edward E. 2002. "An Unprofitable Masculinity." In *Money, Labour, and Land: Approaches to the Economy of Ancient Greece*, edited by Paul Cartledge, Edward E. Cohen, and Lin Foxhall. London: Routledge, pp. 100–112.
- Dandamaev, Muhammad A. 2009. *Slavery in Babylonia; From Nabopolassar to Alexander the Great (626–331 BC)*, revised edition. Translated by Victoria A. Powell. Edited by Marvin A. Powell and David B. Weisberg. DeKalb IL: Northern Illinois University Press.
- Davies, John L. 2009. "The Historiography of Archaic Greece." In *A Companion to Archaic Greece*, edited by Kurt A. Raaflaub and Hans van Wees. Malden MA: Wiley-Blackwell, pp. 3–21.
- Dinsmoor, William Bell. 1950. *The Architecture of Ancient Greece*. London: Batsford.
- Ehrenberg, Ronald G., and Robert S. Smith. 1996. *Modern Labor Economics*, 6th edn. Reading MA: Addison-Wesley.
- Erdkamp, Paul. 2008. "Mobility and Migration in Italy in the Second Century BC." In *People, Land, and Politics: Demographic Developments and the Transformation of Roman Italy 300 BC-AD 14*, *Mnemosyne Supplement* 303, edited by Luuk de Ligt and Simon Northwood. Leiden: Brill, pp. 417–49.
- Eyre, Christopher. 2010. "The Economy: Pharaonic." In *A Companion to Ancient Egypt*, Vol. 1, edited by Alan Lloyd. Malden MA: Wiley-Blackwell, pp. 291–308.
- Fenoaltea, Stefano. 1984. "Slavery and Supervision in Comparative Perspective: A Model." *Journal of Economic History* 44: 635–668.
- Feyel, Christophe. 2006. *Les artisans dans les sanctuaires grecs aux époques classique et hellénistique: A travers la documentation financière en Grèce*. Athens: École française d'Athènes.
- Finley, Moses I. 1980. *Ancient Slavery and Modern Ideology*. New York: Viking.
- Galil, Gershon. 2007. *The Lower Stratum Families in the Neo-Assyrian Period*. Leiden: Brill.
- Garlan, Yvon. 1988. *Slavery in Ancient Greece*, Revised and expanded edition. Translated by Janet Lloyd. Ithaca NY: Cornell University Press.
- Grey, Cam, and Anneliese Parkin. 2003. "Controlling the Urban Mob: The *colonatus perpetuus* of CTh 14.18.1." *Phoenix*, 57: 284–299.
- Gronau, Reuben. 1986. "Household Production: A Survey." In *Handbook of Labor Economics*, Vol. 1, edited by Orley C. Ashenfelter and Richard Layard, 273–304. Amsterdam: North-Holland.
- Hamermesh, Daniel S. 1993. *Labor Demand*. Princeton, NJ: Princeton University Press.
- Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review* 60: 126–142.
- Harris, William V. 1989. *Ancient Literacy*. Cambridge MA: Harvard University Press.

- Hicks, John R. 1932. *The Theory of Wages*. London: Macmillan.
- Hicks, John R. 1963. *The Theory of Wages*. 2nd edn. London: Macmillan.
- Hopkins Keith. 1978. *Conquerors and Slaves: Sociological Studies in Roman History*, Volume 1. Cambridge: Cambridge University Press.
- Hotz, V. Joseph, Jacob Alex Klerman, and Robert J. Willis. 1997. "The Economics of Fertility in Developed Countries." In *Handbook of Population and Family Economics*, Vol. 1A, edited by Mark R. Rosenzweig and Oded Stark. Amsterdam: North-Holland, pp. 275–347.
- Kehoe, Dennis. 2010. "The Economy: Graeco-Roman." In *A Companion to Ancient Egypt*, Vol. 1, edited by Alan Lloyd. Malden, MA: Wiley-Blackwell, pp. 309–325.
- Killingsworth, Mark R. 1983. *Labor Supply*. Cambridge: Cambridge University Press.
- Killingsworth, Mark R., and James J. Heckman. 1986. "Female Labor Supply: A Survey." In *Handbook of Labor Economics*, Vol. 1, edited by Orley C. Ashenfelter and Richard Layard. Amsterdam: North-Holland, pp. 103–204.
- Kossoudji, Sherrie A. 1988. "English Language Ability and the Labor Market Opportunities of Hispanic and East Asian Immigrant Men." *Journal of Labor Economics* 6: 205–228.
- Kuhrt, Amélie. 1995. *The Ancient Near East, c. 3000–330 BC*. London: Routledge.
- Lawrence, A.W. 1983. *Greek Architecture*, Fourth (integrated) ed., revised by R.A. Tomlinson. Harmondsworth: Pelican.
- Lundberg, Shelly, and Robert A. Pollack. 1993. "Separate Spheres Bargaining and the Marriage Market." *Journal of Political Economy* 101: 988–1010.
- Manser, Marilyn, and Murray Brown. 1980. "Marriage and Household Decision-Making: A Bargaining Analysis." *International Economic Review* 21: 31–44.
- Marshall, Alfred. 1920. *Principles of Economics*, 8th edn. London: Macmillan.
- McCloskey, Donald N. 1982. *The Applied Theory of Price*. New York: Macmillan.
- McManus, Walter S. 1990. "Labor Market Effects of Language Enclaves: Hispanic Men in the United States." *Journal of Human Resources* 25: 228–252.
- Oliver, G.J. 2011. "Mobility, Society, and Economy in the Hellenistic Period." In *The Economies of Hellenistic Societies, Third to First Centuries BC*, edited by Zosia H. Archibald, John K. Davies, and Vincent Gabrielsen, 3345–389. Oxford: Oxford University Press.
- Pigou, A.C. 1929. *The Economics of Welfare*, 3rd edn. London: Macmillan.
- Pitt, Mark M., and Mark R. Rosenzweig. 1986. "Agricultural Prices, Food Consumption, and the Health and Productivity of Indonesian Farmers." In *Agricultural Household Models; Extensions, Applications and Policy*, edited by Inderjit Singh, Lyn Squire, and John Strauss. Baltimore MD: Johns Hopkins University Press, pp. 153–182.
- Pitt, Mark M., and Mark R. Rosenzweig. 1990. "Estimating the Behavioral Consequences of Health in a Family Context: The Intrafamily Incidence of Infant Illness in Indonesia." *International Economic Review* 31: 969–989.
- Pitt, Mark M., Mark R. Rosenzweig, and Md. Nazmul Hassan. 1990. "Productivity, Health, and Inequality in the Intra-household Distribution of Food in Low-Income Countries." *American Economic Review* 80: 1139–1156.
- Robertson, John F. 2005. "Social Tensions in the Ancient Near East." In *A Companion to the Ancient Near East*, edited by Daniel C. Snell. Malden MA: Blackwell, pp. 212–226.
- Robinson, Joan. 1933. *The Economics of Imperfect Competition*. London: Macmillan.
- Rosenzweig, Mark R. 1988. "Labor Markets in Low-Income Countries." In *Handbook of Development Economics*, Vol. 1, edited by Hollis B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland, 713–762.
- Saller, Richard. 2012. "Human Capital and Economic Growth." In *The Cambridge Companion to the Roman Economy*, edited by Walter Scheidel. Cambridge: Cambridge University Press, pp. 71–86.
- Sanders, Jimmy M. and Victor Nee. 1996. "Immigrant Self-Employment: The Family as Social Capital and the Value of Human Capital." *American Sociological Review* 61: 231–249.
- Schaeffer, Peter. 1985. "Human Capital Accumulation and Job Mobility." *Journal of Regional Science* 25: 103–114.
- Scheidel, Walter. 2003. "Germs for Rome." In *Rome the Cosmopolis*, edited by Catharine Edwards and Greg Woolf. Cambridge: Cambridge University Press, pp. 158–176.
- Scheidel, Walter. 2004. "Human Mobility in Roman Italy, I: The Free Population." *The Journal of Roman Studies* 94: 1–26.
- Scheidel, Walter. 2005a. "Real Slave Prices and the Relative Cost of Slave Labor in the Greco-Roman World." *Ancient Society* 35: 1–17.
- Scheidel, Walter. 2005b. "Human Mobility in Roman Italy, II: The Slave Population." *Journal of Roman Studies* 95: 64–79.
- Scheidel, Walter. 2008. "The Comparative Economics of Slavery in the Greco-Roman World." In *Slave Systems: Ancient and Modern*, edited by Enrico Dal Lago and Constantina Katsari. Cambridge: Cambridge University Press, pp. 105–126.

- Scheidel, Walter. 2011. "The Roman Slave Supply." In *The Cambridge World History of Slavery*. Volume I. *The Ancient Mediterranean World*, edited by Keith Bradley and Paul Cartledge. Cambridge: Cambridge University Press, pp. 287–310.
- Scheidel, Walter. 2012. "Slavery." In *The Cambridge Companion to the Roman Economy*, edited by Walter Scheidel. Cambridge: Cambridge University Press, pp. 89–113.
- Sicular, Terry. 1986. "Using a Farm-Household Model to Analyze Labor Allocation on a Chinese Collective Farm." In *Agricultural Household Models; Extensions, Applications, and Policy*, edited by Inderjit Singh, Lyn Squire, and John Strauss. Baltimore MD: Johns Hopkins University Press, pp. 277–305.
- Siegel, Bernard J. 1947. *Slavery During the Third Dynasty of Ur*. Memoirs of the American Anthropological Association 66. Menasha, WI: American Anthropological Association.
- Singh, Inderjit, Lyn Squire, and John Strauss. 1986. "The Basic Model: Theory, Empirical Models, and Policy Conclusions." In *Agricultural Household Models; Extensions, Applications, and Policy*, edited by Inderjit Singh, Lyn Squire, and John Strauss. Baltimore MD: Johns Hopkins University Press, pp. 17–47.
- Strauss, John. 1986. "Appendix. The Theory and Comparative Statics of Agricultural Household Models: A General Approach." In *Agricultural Household Models; Extensions, Applications, and Policy*, edited by Inderjit Singh, Lyn Squire, and John Strauss. Baltimore MD: Johns Hopkins University Press, pp. 71–91.
- Sjaastad, Larry A. 1962. "The Costs and Returns of Human Migration." *Journal of Political Economy* 70 (5), Part 2: 80–93.
- Stark, Oded. 1993. "Nonmarket Transfers and Altruism." *European Economic Review* 37: 1413–1424.
- Stark, Oded. 1995. *Altruism and Beyond; An Economic Analysis of Transfers and Exchanges within Families and Groups*. Cambridge: Cambridge University Press.
- Tauchen, Helen V., Ann Dryden Witte, and Sharon K. Long. 1991. "Domestic Violence: A Nonrandom Affair." *International Economic Review* 32: 491–511.
- Thorner, Daniel, B. Kerblay, and R.E.F. Smith, eds. 1966. *A.V. Chayanov: The Theory of Peasant Economy*. Homewood IL: Richard Irwin.
- Todaro, Michael P. 1968. "An Analysis of Industrialization, Employment, and Unemployment in Less Developed Countries." *Yale Economic Essays* 8: 329–402.
- Todaro, Michael P. 1969. "A Model of Labor Migration and Urban Unemployment in Less Developed Countries." *American Economic Review* 59: 138–148.
- Torrence, Robin. 1986. *Production and Exchange of Stone Tools: Prehistoric Obsidian in the Aegean*. Cambridge: Cambridge University Press.
- Trundle, Matthew. 2004. *Greek Mercenaries; From the Late Archaic Period to Alexander*. London: Routledge.
- Tsakirgis, Barbara. 2005. "Living and Working Around the Athenian Agora: A Preliminary Case Study of Three Houses." In *Ancient Greek Houses and Households*, edited by Bradley A. Ault and Lisa C. Nevett. Philadelphia PA: University of Pennsylvania Press, pp. 67–82.
- Turley, David. 2000. *Slavery*. Oxford: Blackwell.
- Warburton, David A. 2005. "Working." In *A Companion to the Ancient Near East*, edited by Daniel Snell. Malden MA: Wiley-Blackwell, pp. 185–198.
- Weeks, Kent R. 2000. *KV 5: A Preliminary Report on the Excavation of the Tomb of the Sons of Ramesses II in the Valley of the Kings*. Cairo: American University Press.
- Westermann, William L. 1955. *The Slave Systems of Greek and Roman Antiquity*. Memoirs of the American Philosophical Society 40. Philadelphia PA: The American Philosophical Society.
- Yuengert, Andrew M. 1985. "Testing Hypotheses of Immigrant Self-Employment." *Journal of Human Resources* 30: 194–204.
- Zhou, Min, and John R. Logan. 1989. "Returns on Human Capital in Ethnic Enclaves: New York City's Chinatown." *American Sociological Review* 54: 809–820.

Suggested Readings

- Becker, Gary S. 1971. *Economic Theory*. New York: Knopf. Chapter 10.
- Ehrenberg, Ronald G., and Robert S. Smith. 1997. *Modern Labor Economics*, 6th edn. Reading MA: Addison-Wesley.
- Killingsworth, Mark R. 1983. *Labor Supply*. Cambridge: Cambridge University Press.
- Nakajima, Chihiro. 1986. *Subjective Equilibrium Theory of the Farm Household*. Amsterdam: Elsevier.

Notes

- 1 Definitions of families and households need not be identical. A household may contain multiple generations of related individuals and parallel groups of mates (husbands and wives) and their offspring, as well as unrelated persons. It could be convenient to define a family as a pair of spouses and their children, and a household as groups of families sharing common cooking and living facilities, with the allowance for the odd, unrelated individual. Slaves would be considered part of a household but not necessarily part of a family. A royal household might contain several hundred unrelated persons, considerably outnumbering the members of the related lineage, but the analysis of such large organizations might be most productively developed separately from more ordinary households. In the present treatment, we will use the terms "household" and "family" more-or-less interchangeably unless we specifically note otherwise.
- 2 We run the risk of using the term "institution" to refer to organizations we otherwise would want to distinguish. It has already been noted that the family or household is the principal labor supply institution, and now that the adjective "institutional" has been used to refer to a large productive enterprise, in distinction to the family production unit, which is still considered "an institution," scope for confusion has emerged. My nomenclatural imagination may be failing me at the moment, for which I beg the reader's forbearance. I believe, however, that reliance on context – the size, composition, and legal status – will be sufficient to clarify meanings about organizations supplying and demanding labor. In the case of an Egyptian temple, we have a case of a large, "institutional" demander of labor – as contrasted with private demanders of labor, most of which were probably small and had no direct affiliation with the government or other official status. The suppliers of labor to a temple would have been individual households, either nuclear or extended families; whether they supplied labor to other, private productive enterprises, including production for themselves, would be an empirical matter.
- 3 Correspondingly, it would repay slave owners to treat trained slaves well, frequently in a fashion superior to the treatment corresponding nonslaves can get for themselves.
- 4 How do we know that indifference curve I_1 is associated with a larger income (or budget) than I_0 ?
- 5 For those who want to understand the time path of hours in Figure 10.15 in more depth, we offer a guide to its derivation. Go back to the first-order conditions for hours (leisure) and consumption described in the previous paragraph. We differentiate those first-order conditions with respect to time – that is, we examine how those formulations respond to the passage of time, in a manner analogous to the way we derived the first-order conditions themselves by seeing how the value of the Lagrangean for the constrained utility-maximization problem varied when we changed the values of hours worked and consumption. That differentiation (the consequence of varying time) gives us two equations, both of which contain the rates of change over time of both hours and consumption. Treating these as simultaneous equations, we solve them for equilibrium values of $\dot{C}$ (which is a notation for $\Delta C_t / \Delta t$) and $\dot{H}$ in terms of the time rates of change in the exogenous variables, p_t and w_t and the two "interest rates" – "the" interest rate r and the subjective rate of time preference ρ . Those solutions for the time profiles of hours and consumption have the forms $\dot{H} = A\dot{w}_t + B(\rho - r) + C\dot{p}_t$ and $\dot{C} = X\dot{w}_t + Y(\rho - r) + Z\dot{p}_t$, in which the coefficients A , B , C , X , Y , and Z represent intricate combinations of marginal utilities and cross-effects between consumption and leisure. When the price of the consumption good is expected not to change, the time paths of C and H are the resultants of the forces of the wage profile and the net effect of the interest rate and rate of time preference. In Figure 10.15, we assume for simplicity that p_t does not change over time. Thus, referring to Figure 10.15, the time profile of the wage is something we could "draw freehand," but once we have done so, the time profile of hours depends on how we drew the wage profile.
- 6 This interpretation may add a new dimension to contemporary thinking about the ancient occupation of mercenary, inasmuch as it posits that a person might have planned a period of such occupation during his lifetime and the plans would have affected his non-mercenary working behavior both before and after his years with the shield. Figure 10.16 says he would have worked less before he went off with the merry band than

he would have if the crew of compatriots had not been an option in the first place; and that he would have worked less when he got back than he would have otherwise also. Trundle (2004, Chapter 2) explores the social setting of fourth-century B.C.E. increase in Greek mercenaries in countries to the east and west of Greece.

- 7 Actually, not entirely independently of when: a period of elevated wages later in the lifecycle will have less effect on λ because it is far in the future, is discounted accordingly, and can contribute less to wealth accumulation because it occurs late rather than early.
- 8 The asset income in this figure can be interpreted two ways. The first interpretation is that while there are no work opportunities outside the household, there are opportunities for exchange of some composite good for money. A variant on this interpretation for a nonmonetized situation is that the composite purchased good measured on the vertical axis is really a combination of different goods, and the household depicted is able to produce more of some of them than it consumes and purchase others. To make good on this interpretation, we would need some tight restrictions on the production functions for the individual goods. The other interpretation retains the autarkic situation of the household and lets the family derive V from productive assets – plows, animals, other tools – but the shape of the production function is not changed (at least in goods-time space of this two-dimensional graph) by the introduction of the capital equipment. A variant on this interpretation relaxes the autarky interpretation of the original situation: the situation with no asset income could still represent production with capital equipment, but some other household owns the equipment and the household depicted pays the marginal product of that equipment in rental for it. The introduction of V amount of asset income is equivalent to giving ownership of the equipment to the household depicted in the figure.
- 9 The treatment of differential substitutability is from Ehrenberg and Smith (1996, 224–226, Figure 7.2).
- 10 The following model is taken from Killingsworth and Heckman (1986, 138–139).
- 11 This exposition is based on Biddle and Hamermesh (1990).
- 12 I have used this subscripting terminology as a shorthand to describe the properties of production and utility functions before. U_i means, in the “delta” notation we have used, $\Delta U/\Delta Z$ or $\Delta U/\Delta T_s$, where the subscript “ i ” refers to either Z or T_s , the interpretation of which is that U increases when Z increases (because in the text we have said that this quantity is positive). The “double i ” subscript indicates the doubled delta – the change in the rate of change: for example, $\Delta(\Delta U/\Delta Z)/\Delta Z$, which, of course, characterizes the degree of curvature of the utility function. These quantities are ordinarily negative, meaning that, while U increases when we have more Z -goods, as we get still more Z -goods, the increase in utility gets smaller. The “ ij ” subscript notation is the cross-effect on the marginal utility of either Z -goods or sleep as more of the other item is had: in the delta notation $\Delta(\Delta U/\Delta Z)/\Delta T_s$. These quantities are usually positive, meaning that the marginal utility of, say, Z , is enhanced if we also have more sleep. I use this notation for its compactness, not for obscurity.
- 13 The term D at the end of this expression is the determinant of the matrix of coefficients derived from the first-order conditions for maximizing utility, which result in equations that look much like the numerator and denominator of the ratio of marginal costs of Z goods and sleep. To reach the expression for $\Delta T_s/\Delta V$ shown in the text, those first-order conditions are further “differentiated totally” to see how they change as each of the variables, both endogenous and exogenous, change. “Total differentiation” is pretty much the same procedure as varying the magnitudes of the variables in a constrained maximization problem to get the first-order conditions, which tell us what level of just one of the variables will give the largest (or smallest, whichever we’re looking for) value of the expression the economic agent is trying to maximize (or minimize); the principal difference is that with total differentiation, we perform those variations in the magnitudes of each of the variables, so that the entire expression is in terms of differences in the variables, rather than in their levels. (For example, an expression in levels would look like $a + b = c$; the same expression in differences would be $\Delta a + \Delta b = \Delta c$; we’ve just totally differentiated the expression $a + b = c$.) The resulting equations are put into matrix form, one matrix of which is the coefficients on the endogenous variables, T_s and Z . The determinant of this matrix, represented by D , is a single number derived by a number of multiplications of the individual terms in the matrix. The exponent -1 used on D in the expression for $\Delta T_s/\Delta V$ indicates that the other terms are divided by D , which indicates that the system of simultaneous equations that, when solved, yields expressions for $\Delta T_s/\Delta V$, $\Delta T_s/\Delta w_1$, and $\Delta T_s/\Delta w_2$, was actually solved by a procedure known as Cramer’s rule, which

involves dividing one determinant by the determinant of the coefficients matrix of the endogenous variables. This may be more than some readers wanted to know, but it at least shows that there is no mystery about how the expressions in the text are derived from the original constrained maximization problem.

- 14 For the interpretation of our use of this subscripting notation to substitute for the clumsier notation using deltas to indicate positive and decreasing marginal products, see note 12.
- 15 Exactly, the slope of the isoquant is the ratio of the marginal products of capital and labor, which must be equalized to the ratio of their marginal costs, or the slope of the relative factor price line. With one factor in fixed supply, there's only one instrument at the producer's disposal to maximize his profit, and it's not necessarily the case that his short-run profit maximizing-input quantities will be efficient (that is, involve a tangency). In the long run, when both factors are variable, his profit maximizing input choice will also be efficient necessarily. In the short run, the question is whether he can hire a large or small enough quantity of labor to get an exact tangency to one of the isoquants in Figure 10.26. If the family of isoquants is continuous in the capital-labor space of that diagram, his chances are better than if the technology restricts him to discrete quantities of labor that must be used in conjunction with K capital.
- 16 Some authors measure the elasticity of substitution as $\% \Delta(L/K) / \% \Delta(w/r)$, in which case its sign is ≤ 0 . Take care to note how it is being defined in any particular application.
- 17 Although moving along a given isoquant is remaining within the same technological choice set, choices of different points along an isoquant are themselves different choices of technologies, and such changes along an isoquant will involve costly reorganizations. The term "technological change" more commonly refers to changes in the shapes or positions of isoquants. Be careful to distinguish between the meanings of the two terms.
- 18 Nothing is really lost by adopting this convention: authors have simply written in terms of one parameter being "more elastic" or "less elastic" as some other parameter increases in value, and as long as the signs are all reversed consistently, sense emerges. The practice is virtually uncommented upon in economics textbooks that discuss derived demand elasticity, and the practice goes back to the original treatment of Alfred Marshall, who reversed the signs on both demand elasticities, to make them positive in his mathematical

exposition (Marshall 1920, Book V, Chapter VI, 316–326, and Mathematical Appendix, note XV, 701–702). Another well-known treatment, by A.C. Pigou, remained a verbal exposition but spoke in terms of all positive elasticities (Pigou 1929, Part IV, Chapter V, 679–688). The definitive derivation, by J.R. Hicks in 1932, explicitly reversed the signs of the product demand elasticity and consequently measured the derived demand elasticity positively (Hicks 1932, 242–246; 1963, 373–378). His analysis was followed a few years later by Joan Robinson's examination of derived cross-elasticities, also measured positively (Robinson 1933, 259). Interpreting the results is probably easier when all the elasticities are measured positively, which may account for this convention that has crept into the standard treatment, even in contemporary textbooks. Anyway, we consider that it would cause more confusion for readers who may read the standard treatment elsewhere to reverse the signs of the demand elasticities in the current treatment, so we bow to convention. Nonetheless, the reader should be aware of the sign switch and not be misled by reading into it anything other than a convention. The theory of derived demand really does rely on quantities demanded of both products and factors falling when their own prices rise.

- 19 To the extent that it's really sensible to talk about derived demand for labor at the level of product aggregation that would be associated with an entire labor market.
- 20 The general form of the inverse cross-elasticity (of the price of the other factor with respect to the availability of labor) is $1/\mu' = [s_L(\eta - \sigma)] / \{\eta\sigma + \varepsilon[s_L(\eta - \sigma) + 1]\}$.
- 21 Factor mobility is an important concept, which need not refer only to geographical mobility of factors of production. In general, factor mobility refers to the responsiveness of the specific employment of an input to factor price differentials. Most inputs other than labor are owned by someone, so the concept refers to the ability of a factor owner to change to whom he or she sells or rents units of a productive input in response to different price offers. With labor, except slave labor of course, the owner of the factor is the factor itself, and factor mobility refers to a person's ability to offer his or her services to the highest bidder – accepting, of course, that factors other than the wage alone may affect the utility of an employment offer. Factor mobility exists in degrees. Zero factor mobility means that inputs are absolutely stuck where they are currently used; if a higher price is offered by producer 2 than

producer 1 is now paying for factor A, no units of factor A would be reallocated by their owner from producer 1 to producer 2. Perfect factor mobility means that employment responds instantaneously and fully to price differentials, so that no such price differentials can remain, again allowing for side-conditions of factor employment that must be included along with price in assessing the utility to a factor owner of where he rents or sells his inputs. In between these two extremes is a lot of room for interesting reasons for partial factor mobility – uncertainty of lots of things, including information; contracts that are costly to break; missing markets; “side-deals” such as intrafamily exchanges. Most basic expositions of price theory begin with perfect factor mobility as a benchmark because so many circumstances can lead to degrees of imperfect mobility, especially in the short run, that no general understanding of any allocative processes would be possible unless those particularities were temporarily ruled out of the analysis. They can be brought back in once the mechanics of the allocational benchmark are understood, and their impacts thus be better determined.

22 Modified from an example of seventeenth and eighteenth century C.E. Britain, in McCloskey (1982, 532–533).

23 A formulation contributed by Sjaastad (1962), which placed migration decisions squarely within human capital decisions. A major contribution by Schaeffer (1985, especially sections 2 and 3) extended Sjaastad’s human capital framework to lifetime migration programs, showing that modeling migration decisions as one-time, permanent choices could recommend moves that would be suboptimal in a multi-move plan. The idea that people plan out their entire lifetime’s mobility and location decisions in advance, when they are young and callow seems to stretch credibility but a moment’s reflection suggests that people may, at early ages, plan on returning to their family homes after a working life elsewhere, possibly even with a move or two thought about during the working life. If one considers Schaeffer’s lifecycle program flexibly, it is not that farfetched.

Both Archibald (2011) and Oliver (2011) raise issues of lifetime migration patterns, mixed, however, with issues of occupationally created geographical mobility, as in Archibald’s discussion of Feyel’s (2006) analysis of construction workers of various skills at Delos, actually the equivalent of contemporary oilfield workers who live in Louisiana and Houston and spend months at a time, years on end, working at sites from Egypt to

the North Sea; and socio-economic mobility at a fixed location, raised by Oliver as an alternative to geographical mobility. Oliver’s discussion, in particular, raises the theoretical issue in the context of, say, Schaeffer’s model of lifetime migration, of changing opportunities at many or all of the possible migration destinations over a person’s lifetime, none of them being predictable. How could anyone at an initial location plan even the final location, much less the intermediate stops, in such a migration plan? I’d suggest that such a case simply provides an opportunity for the student of the ancient migration / mobility topics to think about ancient decision makers in terms of very diffuse expectations and stochastic opportunities. People adapt to confusing and unpredictable times, but planning horizons probably contract considerably. High interest rates (a signal of great uncertainty) can make possibilities several years into the future irrelevant to current decisions.

24 In an analysis of rural-to-urban migration in contemporary developing countries, Harris and Todaro (1970) posed the problem of the urban wage’s attraction to migrants thus: rural workers would continue to migrate to cities as long as the urban wage times 1 minus the urban unemployment rate equaled the rural wage. This formulation has become the foundation for much subsequent research on labor market equilibration in a variety of settings. It is but a step from that formulation to deflating nominal wages at origins and potential destinations by local costs of living, which themselves may be endogenous, as well as accounting for employment probabilities.

25 Archibald (2011, 50) makes reference to the human capital contained in social networks, which could help immigrants, although she does not pursue the theme in the context of migration, using it instead in her thoughts on innovation (56–61). Economists and sociologists have studied the effects of ethnic enclaves on the economic progress of international immigrants as human capital issues: Yuengert (1985); Kossoudji (1988); Zhou and Logan (1989); McManus (1990); Sanders and Nee (1996); Borjas (1998, 2000).

26 Note that we say “owners of other factors in the region.” These owners need not reside in the region; we need only that their factors are used and produce income initially accounted in the region.

27 This conclusion holds constant one factor that may have been quite important in many ancient migrations, particularly those of larger scale: what do the immigrants do for land? If the land in the receiving country is fully owned but incompletely

- occupied and the immigrants occupy parts of it, natives take an immediate reduction in their effective ownership of land. If immigrants, who let's assume are predominantly agriculturalists as are most of the natives, initially take up occupations as landless farm laborers under these aggregate cost conditions, native land owners will feel little initial effect. Land rents will go up by some amount that will be small relative to total income. However, as immigrants go through their lifecycle, and into their next generations, they may accumulate wealth which they invest in land. If the total quantity of improved land is fixed, they will be buying it from natives, whose wealth will rise as a consequence of the purchases. The rise in immigrant land ownership will cause a secondary redistribution of income in excess of area *B* in Figure 10.36 even though the positions of the curves and the magnitudes of the areas marked out in that figure will not change. Such orderly redistribution of land ownership could have affected popular sentiment toward immigrants. Less orderly land redistribution, say through squatting on ineffectively utilized land that migrants nonetheless considered theirs, would have been an even more likely cause for concern.
- 28 The original external publication is Todaro (1969), developed from his dissertation, Todaro (1968). It was followed by a version explicitly modeling rural and urban production: Harris and Todaro (1970).
 - 29 The joint incentive is that if one of the family members suffers a major income failure, the other family members are likely to suffer as well, so it pays them all to insure one another's activities by, say, taking up the slack a sick family member temporarily leaves in the family pull-rope.
 - 30 This is not just a matter of the husband "giving" the wife an arbitrary amount of *Z* goods in income. This quantity is the amount of *Z* goods that the wife actually produces in the marriage. We are dividing up factor income between wife and husband just like we have done in the theory of the firm between labor and land inputs.
 - 31 Some – but not all – writers consider the concept of peasantry to be anachronistic for the ancient Mediterranean region – or ancient times in general.
 - 32 The "reasonable circumstances" being a woman's relative marginal productivity twice that of the husband's, nonincreasing returns to scale in household production, and a wealth elasticity of 1; see Becker (1991, 93–94).
 - 33 The implicit bargain in the unitary altruistic model is the decision to follow one individual's (possibly altruistic) preference function. See Manser and Brown (1980).
 - 34 The most important distinction between cooperative and noncooperative game theory is that direct communication is not possible in the latter but is in the former. Accordingly, in cooperative bargaining, the parties are able to make joint, binding agreements. The implication for the analysis of marriage is that the cooperative approach supposes that spouses can make such binding agreements within, as contrasted to before, the marriage. The noncooperative approach takes as its starting point that such agreements have broken down – that is, that they aren't possible.
 - 35 But Bergstrom (1996, 1926–1929) characterizes a model of repeated disputes within a marriage, with smaller penalties for disagreements than divorce (generally marital dissolution) – what he calls "burnt toast."
 - 36 As Peter Schickele of PDQ Bach has expressed so pithily, "It's true that the best things in life are free, but the really good things cost a lot of money." The transferrable utility model focuses on a utility frontier that two (or more, depending on institutional possibilities) spouses create through their production activities, essentially like a production possibilities frontier, which shows the maximum amounts of two goods that can be produced from a given set of resources with a given technology. There may be one good, with properties in a special type of utility function, possessed by each spouse, which lets it be varied in amount while leaving the marginal utility relations among all other consumption goods unaffected, thereby transferring utility from one partner to the other. Such transfers can permit spouses to generate utility combinations on their utility possibilities frontier rather than being confined to inefficient points inside the frontier because of the distributions of endowments between the two partners.
 - 37 The following treatment relies on Behrman (1997, 129–140); also see Behrman *et al.* (1982).
 - 38 See also Pitt and Rosenzweig (1986) and Behrman (1997, 154–167).
 - 39 The reader may wonder about the empirical implementation of the concept of a person's endowment of health: how could you ever observe and measure it? Pitt *et al.* (1990, 1145–1147) used an indirect technique on detailed, individual observations of physical characteristics such as mid-arm circumference, skinfold thickness, height, weight, dietary information, and calorie output. With these data, they estimated the health production function just

- described and extracted as a residual measure, an index of health endowment.
- 40 The following exposition is adapted from Birdsall (1988, 503–505), Becker (1991, Chapter 5), and Hotz *et al.* (1997, 292–308).
 - 41 These phenomena may remind some readers of the Chayanov model of the Russian peasant farm, popular in anthropological studies since the 1950s at least: Chayanov (1925) in Russian, and in English, Thorner, Kerblay, and Smith (1966). What I am going to present is indeed basically a formalization of the Chayanov model.
 - 42 I use an example devised by Hamermesh in (1993, Chapter 10).
 - 43 I adapt the following exposition from Strauss (1986). Also see Barnum and Squire (1979, Chapter 3); Singh *et al.* (1986) and, for a wider overview, Rosenzweig (1988).
 - 44 The budget constraint also could be expressed as $p_m X_m = p_a (Q_a - X_a) - w(L - F)$, in which L is the farm's total demand for labor and F is labor supplied by the family, with $T = X_L + F$. If $L - F > 0$, the family hires labor in, and the farm has to "finance" its labor hiring through its marketed surplus $Q_a - X_a$. If $L - F < 0$, the family supplies labor off-farm and brings in gross labor income as part of its "cash" budget.
 - 45 Strictly speaking, this "non-wage" income *could* be wage income, such as would come from remittances from a household member who migrated to a city and sends some of his or her labor income back to the family at home. The distinction we're making here is between income that the people remaining on the farm produce with their own time and income that comes from some other source that does not take time away from household or farm activities. Remittances, even if they are wage income, derive from labor that has already been taken away from production possibilities on the farm.
 - 46 Note that some of the outputs subscripted with j here refer to some of the consumption items subscripted with i in the budget constraint; we're just using different "counters"; thus when the output we refer to as "output j " is good a , p_j is p_a and when the consumption good we discuss is good a , p_i from the budget constraint is p_a : $p_j \equiv p_i \equiv p_a$.
 - 47 Notice that the profit function contains as arguments output prices as well as factor prices. While varying the factor prices in a profit function yields the demand for the corresponding input, varying the output prices yields the corresponding supply functions.
 - 48 Basically, by virtue of our knowledge of the way indifference curves "bend" in commodity space and the way production isoquants bend in input space. These shape properties translate into how the expenditure and profit functions "bend" in their respective commodity and input price spaces. (By "spaces" I refer to a two-dimensional graph with one consumption good on the vertical axis and another one on the horizontal axis, or a two-dimensional graph with labor inputs on one axis and land inputs on the other.)
 - 49 e_L is the demand function for leisure; e_{LM} characterizes how the demand for leisure changes when the price of the purchased good changes. If leisure and the purchased good are substitutes, an increase in p_m pushes consumers out of X_m and into leisure, effectively increasing the demand for leisure. In terms of the expenditure function, if p_m rises, to keep utility constant, "expenditures" on leisure will rise and those on X_m fall. Thus, $e_{LM} > 0$.
 - 50 This characterization of quotas is more general than may first appear. One might wonder about the imposition of a quota on production rather than on sales. But if the household doesn't sell what it produces to the agency imposing the quota, how much it produces is immaterial to that agency; sales are the concern of the agency.
 - 51 For an overview, see Bell (1988). On tying between tenancy contracts and other markets, including the market for effort, see Rosenzweig's (1988) review of various studies.
 - 52 $\text{Wage}^t = \text{price}$, where r is the discount or interest rate and t is years.
 - 53 The Solonian emancipation of the thetes in the early sixth-century B.C.E. is cited commonly as the source of a corresponding "labor shortage" or "scarcity" in Greek agriculture, which initiated the increase in scale of Greek chattel slavery: Finley (1980, 89–90; Burford, 1993, 209). Why emancipation of thetes in Attica would lead to the extensive chattel slavery recorded on Chios and Aegina at the same time or even earlier (Burford, 1993, 209; Scheidel 2008, 118) is an open question, of course. Davies (2009, 17) attributes the expansion of Greek chattel slavery "beyond the domestic context visible in Homer" late in the Archaic period as an indirect consequence of the recovery and enlargement of craft and technological skills lost during the Greek Dark Age, because "only a bought labor force could be moved forcibly to locations, whether mines or workshops or households, where extra labor was needed." One thinks of some of the large construction projects of the late Archaic period, at least for the heavy lifting if less so for more skilled activities

- (Dinsmoor 1950, Chapters 2–3; Lawrence 1983, Chapter 10).
- 54 Finley (1980, 89) falls back on “psychology,” which amounts to the same as tastes.
 - 55 Kuhrt (1995, 62) notes that during the Ur III period in Mesopotamia, 40% of the slaves appearing in sale documents were indigenous – poor families selling their children into slavery. She even cites reports of sons selling their mothers into slavery. Galil (2007, 192) reports similar practices in the Neo-Assyrian period in northern Mesopotamia.
 - 56 Garland (1988, 46–54) cites the sources of slaves into Classical Greece as prisoners of war, piratical captures and in the cases of interhellenic wars, fellow Greeks – but from other cities. Scheidel (2005b) employs detailed, alternative demographic calculations to gauge the contributions of imports and natural reproduction to the slave population of Roman Italy from the Early Republic to the Early Principate. Kuhrt (1995, 534) cites prisoners of war as the predominant source of slaves in Mesopotamia during the Neo-Assyrian period.
 - 57 The appeal to lower labor turnover costs with slave labor than with free labor surely was a tertiary motivation at best for holding slaves (Scheidel 2008, 111; 2012, 102 citing “transaction costs,” apparently turnover costs again, 203). Compared to losses of free and slave labor to adult mortality, turnover costs of free labor – taking other jobs where? – may not even have been observable to Roman employers.
 - 58 Galil (2007, 194) reports the common occurrence of guarantee clauses in slave sales, which protected the purchaser to the loss of a slave to particular diseases within 100 days of sale and from “crime” forever. The former is easy to understand, the latter less so.
 - 59 Westermann (1955, Chapter IV) offers such a background of incentives provided to slaves in Classical Attica by both legal institutions and public practice. His other chapters are replete with insights into treatment and incentives.

Land and Location

11.1 The Special Characteristics of Land

Land is one type of capital, like equipment or an education, but it has some distinguishing features that warrant separate treatment as one of the three principal types of factors of production. First, it is fixed in location, unlike labor or capital. Frequently, it is in relatively inelastic supply, which means that owners commonly can receive more for its use than the minimum that would be required. Land is strikingly heterogeneous, differing as it can, even within a small region, in soil characteristics, topography, microclimate, and the natural or cultivated vegetation on it. In many agrarian societies, land may be the principal form of wealth and, as such, provides the primary source of political as well as economic power.

It has long been common to describe land as the “original and indestructible” factor of production – one that can be neither produced nor destroyed. Both contentions are false. Effective units of land can be created by such investments as terracing and fertilization. Land can be effectively destroyed through overuse, which leads to erosion, salt accumulation, and other problems (Jacobsen and Adams 1958; Adams 1981, 151–152). Continuous maintenance of

land is required for it to retain its productivity at a given level.

Land is a particularly prominent factor of production in agriculture, and agriculture claims a large share of a society’s stock of land in use. Land tenure, the contractual conditions under which land is owned and operated, was alluded to briefly in Chapter 7, although it is a subject that warrants considerably more detailed attention than can be devoted to it here. First, the subject matter commonly called “land tenure” is described more accurately as “farm tenure,” and second, frequently there are interlinkages between farm tenure and other agricultural and rural allocation mechanisms that make discussion in a single place more coherent.

This initial treatment abstracts from the locational characteristics of land as a factor of production. Pursuing the fixed location of land, we will introduce three major components of the economic theory of location in this chapter. The first branch of location theory we introduce, in section 11.3, deals with the different uses of land in different locations relative to some fixed site, possibly a town or market place. This is commonly called agricultural location theory, but the same model forms the basis for the economic analysis of the spatial structure of cities, addressed in Chapter 12. The allocative mechanisms of this

model are quite general and can be said to form the basic mechanism of all of location theory. Section 11.4 focuses on the location of a production facility – a manufacturing plant. We begin with the location of an individual plant then consider the interaction within a region of different plants that produce the same good or goods. The final section deals with the locational aspects of consumption: how far will people travel to obtain products and services. While this theory may look rather abstract, it adds considerable realism to the theory of consumption as we treated it in Chapter 3. The subsection on hierarchies of market places may be familiar to some readers as central place theory.

As transportation is so intricately intertwined with the use of land and the interconnections of locations, this is a reasonable place to pull together some of the resource allocation issues associated with that activity. Section 11.6 divides transportation issues into infrastructure – roads, bridges, and so forth – and movable equipment – wagons, ships, and so forth. A closing subsection focuses on the pricing of transportation services.

11.2 Land as a Factor of Production

This short section provides a benchmark against which to view the more locationally oriented sections to follow and to compare the special features of labor and capital developed previously.

11.2.1 Supply

Abstracting from locational differences among units of land, the supply curve of land is the annual cost of bringing a marginal unit of land into use. Supply cost includes the interest payment on the original investment required to clear or otherwise bring the land into use, including terracing, irrigation, and so forth. Additionally, the supply cost includes charges to pay for maintenance activities to keep the land at its current level of physical productivity.

Figure 11.1 shows a supply curve of land intersected at point B by a value-of-marginal-product curve, or the demand curve for land, in a particular region. The gross rent on each unit of land

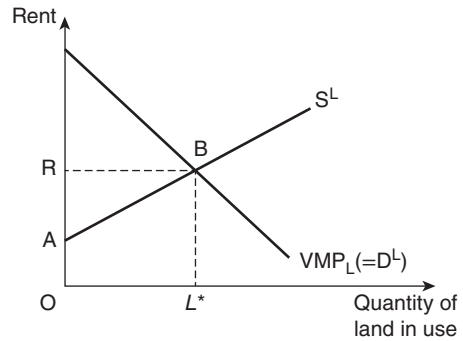


Figure 11.1 Supply and demand curves for land.

is vertical distance OR , and L^* units of land will be used. Following the supply curve S^L out to the right beyond point B indicates that there is more land available in the region than people choose to use during this year. The triangular area RAB is the annual net rent accruing to land owners in the region. The area under the supply curve, $OABL^*$, is the resource cost of using L^* land during this year.

A change in the interest rate will shift the intercept of the supply curve. The technology of clearing new land will affect the supply curve's slope. Maintenance costs will shift the intercept. Alternative demands for land, such as urban uses, also will alter the slope.

11.2.2 Demand

The previous sentence just altered the terms in which we must think about the demand for land. Different activities use land, and they all possess their own demand curves for land. It is easy to think only of agriculture as the source of the demand for land, and in thinking about many Mediterranean societies in antiquity, we probably wouldn't be far wrong. Nevertheless, as we move from thinking about larger regions, say all of southern Mesopotamia, to smaller regions, say a 50-square-mile area around Ur, the alternative demands for land become more important in any analysis we might undertake in which land is subject to allocative actions. Cities, rural housing, irrigation systems to some extent, even cemeteries, exert demands for land.

Having issued this warning, let's focus on agriculture's demand for land. Working from a production function for agricultural products, through the first-order conditions for optimal resource allocation, we arrive at a demand function for land (its value-of-marginal-product function) of the form $L^d = L_d(r, w; p_a)$, where r is the rental on land, w is the wage rate, and p_a is the price of the agricultural product (you can think of a price index of all agricultural products). As r increases, the demand for land decreases (we work back up the VMP_L , or demand, curve; it doesn't shift down). As the wage increases, the VMP_L curve falls, assuming that land and labor are gross substitutes, as we discussed that term in the previous chapter. An increase in the agricultural product price will shift the entire VMP_L curve upward.

Actions that make land more productive will increase the demand for it – shifting the demand curve upward: increasing the use of fertilizer, using more labor, providing a timely supply of water (possibly through irrigation). More people making offers for land will shift the aggregate demand curve for land to the right. With a fixed supply curve of land, this will not increase the quantity of land used but will raise the rent on land. Population growth in a fixed territory would fit this description.

11.3 The Location of Land Uses

The first of the three principal bodies of location theory addresses how land at any particular location, or zone of locations, is used. How land is used has many dimensions: what is grown on or otherwise done with the land; ratios of inputs, both to land and to one another; techniques of production; contractual agreements; personal characteristics of people who live or work at various locations; and so forth. The first subsection introduces the Thünen model of land uses and land pricing in an agricultural setting, and the second sub-section shows how the bid-rent function of the basic Thünen model can be used to address a diversity of location problems involving the use of continuous land areas.

11.3.1 The Thünen model¹

The Thünen model drastically simplifies a landscape – typically a rural landscape but increasingly commonly an urban landscape – by abstracting from all topographical variations, yielding what is commonly called the “transportation surface” or “transport plain,” with identical transportation costs in all directions, to begin the analysis of the motivations for conducting activities at different locations relative to some central location. To study locational motivations in particular places, the student would want to account for such variations, but to gain insights into the economics of location, the simplifications of terrain are enormously useful: why deliberately conflate the effect of an uphill slope with a transportation price?

Let's start out with a large plain that is undifferentiated topographically. Without worrying a lot at the moment about the boundaries of this plain, let's suppose also that toward the center of the plain is a town. Only one crop can be grown in the region, and the full crop from each unit of land is sent to the town for sale. Once the crop from a distant farm gets to the town, each unit of the crop gets the price p_Q , but out of that revenue, the farmer must pay the transportation cost of t per unit of crop per unit distance (call it a mile). Let's label the units of distance “ k .” If some units of this crop are hauled k miles, the transportation cost on each unit for that distance hauled is tk . If we call p_Q the market price of the crop, then $p_Q - tk$ is the farm-gate price at a site located k miles from the town. Figure 11.2 shows the market and farm-gate prices as we move away from the town in only one direction. At location k_1 we subtract tk_1 from p_Q to get the farm-gate price $p_Q - tk_1$ at that site, and similarly for site k_2 . The diagonal line from p_Q at the market is the farm-gate price line.

Now turn to the production technology. It's simplest to start out with a fixed-coefficients farming technology that uses land and labor. Assume that the wage rate is the same throughout the region. Then at each location as we move away from the town, the non-land production cost of the crop is the same; call it “ a ” per unit of crop. So the costs that a producer at location k has

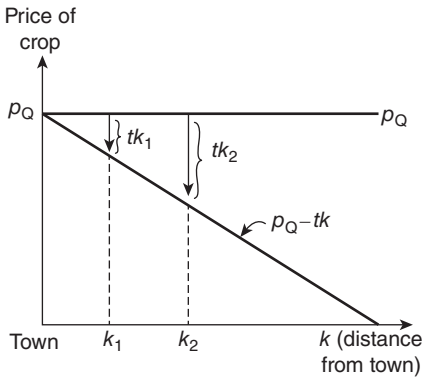


Figure 11.2 Transportation costs and the farmgate price of a crop.

to pay on each unit of the crop he produces there are $tk + a$. With a fixed-coefficients production function, producers produce exactly the same quantity of crop on each unit of land (called the yield per acre, Y), so each unit of land yields a net revenue of $(p_Q - tk - a)Y$. Figure 11.3 draws these values, giving a value of 1 to Y , the yield per acre. The horizontal line at a measures the vertical distance of non-land production cost at each location. Measuring down from p_Q gives us transportation costs at each location. These two costs do not entirely absorb the revenue from selling the crop at the town until we reach location k^* , which defines the edge of land use by virtue of the fact that if someone were to produce this crop at a location further from the town than

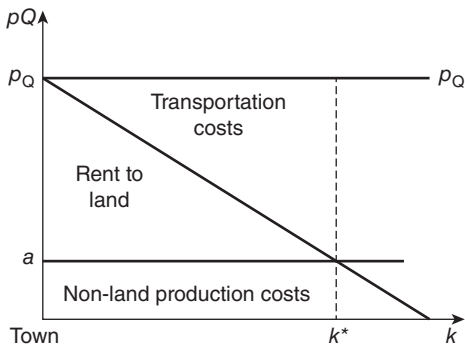


Figure 11.3 Transportation costs and land rent.

k^* , he would receive a negative revenue – he’d have to pay to produce it. As Figure 11.3 indicates, this positive net revenue is rent to land at each location. More specifically, it is the rent to the location since there is nothing except location to distinguish one unit of land from another so far, according to our stringent assumptions. The only supply cost associated with a unit of land is the cost of getting the crop from the site to the market in town. Then the rent to a unit of land at location k is $R(k) = (p_Q - tk - a)Y$. Only when the yield per acre is defined so as to be 1 can we label the axis of a diagram like Figure 11.3 with just the market price of the crop; measuring rent on the vertical axis requires that we measure the product of market price and yield per acre.

Next let’s introduce a second crop; call it Z , and subscript the corresponding prices and costs with “ Z ”: $R_Z(k) = (p_Z - t_Z k - a_Z)Y_Z$. Notice that by distinguishing the transportation cost of crop Z from that of crop Q , we’re implying that differential transportation cost will be one source of differential rent-generating capacity between these two crops. In Figure 11.4, to simplify the drawing, we let $a_Q = a_Z$. At locations between the town and $\bar{k}$, crop Q offers producers larger rents on land. At $\bar{k}$, either crop will deliver the same rent to land so producers will be indifferent between producing the two crops at that location. From $\bar{k}$ to k_Q^* , crop Z can offer more rent to land than crop Q . In fact, beyond k_Q^* , crop Q offers crop Z no competition at all for land. The vertical dashed line at location k_i in Figure 11.4

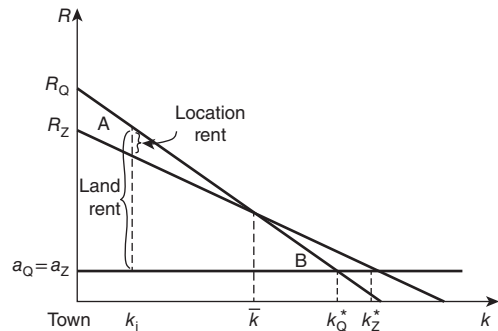


Figure 11.4 Location rent and land rent.

distinguishes between the full rent to land at this location and the location rent accruing to crop Q. The location rent is the surplus that one crop can earn at a location above what the best alternative use of that location (growing crop Z in the present case) can yield. Location rents decrease steadily from the town to the land-use switch point $\bar{k}$, because the next-best alternative is getting relatively better with distance. Why is this happening? The only possible explanation for the shallower slope of the rent line for crop Z is a lower transportation rate, t_Z : the transportation cost (not the transportation rate) is the only variable that changes with location. From $\bar{k}$ to k_Q^* , the location beyond which it would not pay people to produce crop Q, the location rent accruing to crop Z grows larger because the next best alternative to growing crop Z (that is, growing crop Q) is getting worse. Between k_Q^* and k_Z^* , the location rent on crop Q falls to zero because its alternative, which is growing nothing, offers a constant rent of zero, while the absolute surplus from growing crop Z falls continuously, as it would do all the way from the town. The land rent throughout this region is now the outer envelope of the R_Q and R_Z lines. Area A is locational rent to crop Q, and area B is the locational rent to crop Z. The full land rent to either crop is the entire vertical distance between the highest rent line and the horizontal line $a_Q = a_Z$.

One thing we can't tell from this diagram is the combination of market price per unit of crop and yield per acre that gives one crop the advantage in being able to afford interior production locations. If we knew that $p_Q = p_Z$, then we'd know that $Y_Q > Y_Z$; conversely, if we knew that the yields were equal, we'd know that $p_Q > p_Z$. However, with equal transportation rates, $t_Q = t_Z$, a higher yield per acre would impart a steeper slope to the corresponding crop's rent curve because the transportation costs are being expended on a larger quantity of crop at any given distance. Consequently, just from looking at this rent diagram, we can't tell which crop has the higher transportation rate. It's not impossible that we'd have the combination $p_Q < p_Z$, $Y_Q > Y_Z$, and $t_Q > t_Z$.

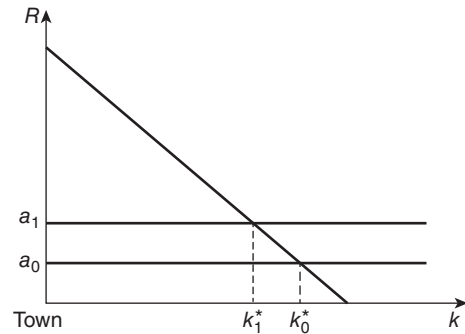


Figure 11.5 Responses of the edge of land use to non-land production costs.

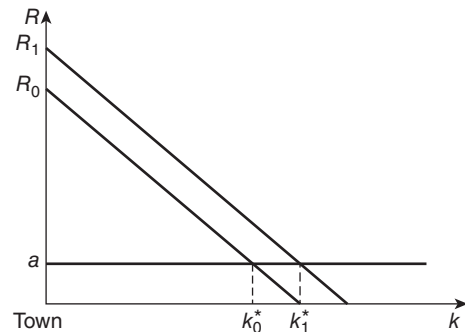


Figure 11.6 Responses of the edge of land use to market price of crop.

Let's return to the one-crop case for a moment to see how the edge of land use changes in response to the key variables in the problem. Suppose that non-land costs (assumed to be just wage costs here, but they could come from a number of sources) increase from a_0 to a_1 as in Figure 11.5. The edge of land use shrinks from k_0^* to k_1^* . Rent falls on all units of land. Suppose that the market price is higher. Figure 11.6 shows the rent line moves up parallel to the original rent line and the edge of land use moves outward from k_0^* to k_1^* . Rent increases on all previously used units of land and becomes positive on locations between k_0^* and k_1^* that previously were uneconomic to cultivate. Next, suppose that yield increases. This one is a little trickier, as Figure 11.7 shows: the rent line pivots around the non-land cost line, a ,

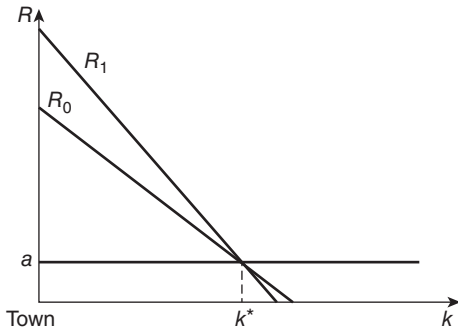


Figure 11.7 Responses of the edge of land use to yield.

so as to keep k^* constant. To see that this happens, write out the expression for rent at the edge of land use: $R(k^*) = (p_Q - tk^* - a)Y = 0$, since rent is zero at the edge of land use. Rearrange that expression to get $(p_Q - a)Y = tk^*Y$, and Y cancels, leaving $k^* = (p_Q - a)/t$. Since the expression for yield, Y , does not enter the formula for k^* , yield does not affect the distance at which the edge of land use occurs. It does, however, make for a steeper rent curve. This result depends critically on the assumption that the market price of the crop is given (“determined by forces outside the region” is the only reasonable interpretation of this assumption), which directs our attention to that assumption.

In many of the applications that readers will have in mind, it is reasonable to suspect that the quantity of a crop produced in a region would affect its market price. (This can happen even if only a portion of the crop is shipped off farm for external sale, that is, even if we’re shipping only the “marketed surplus” to the town.) In addition to the formulation for rent, we need to specify the demand conditions for our crop. Staying with the one-crop case for the moment, suppose that its market price in the town is determined by the intersection of demand and supply curves. The supply curve of the crop is simple to determine with the fixed coefficients production assumption: Y units of the crop are produced at each location. To increase the quantity supplied we must bring more land into production. With the

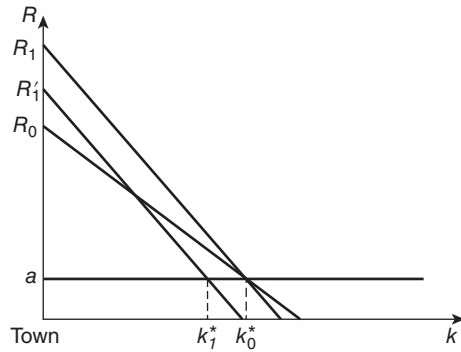


Figure 11.8 Land use response to a yield increase that affects crop price.

constant non-land cost of production, the only source of rising production (supply) cost is the larger transportation cost of getting additional supplies to the town. In Figure 11.8, the increase in yield per acre has the initial effect of pivoting the rent line around location k_0^* , from R_0 to R_1 , but the ensuing outward shift in the supply curve for the region causes the market price to slide down the constant demand curve (let’s assume that the income effect of the increased yield on the demand for the crop is sufficiently small to be ignored). The lower price times the higher yield drops the new rent line down parallel to R_1 , such as R'_1 . The edge of land use moves toward the town with the reduction in the market price, to k_1^* . Intuitively, this is quite reasonable: With only the price effect having any discernible influence on demand, when more of the crop can be grown on a unit of land, a smaller area will suffice to produce the crop. The quantity of the crop represented by the rent line R'_1 is greater than the quantity represented by rent line R_0 but less than that associated with rent line R_1 .

Consider, next, two crops with endogenously determined market prices – that is, with their local prices determined by the interaction of local demands with local supplies. Call the two crops A and B, with crop A able to pay the differential rent on the inner ring near the town and crop B better suited to pay the differential rent in the outer ring of cultivation. Consider an exogenous

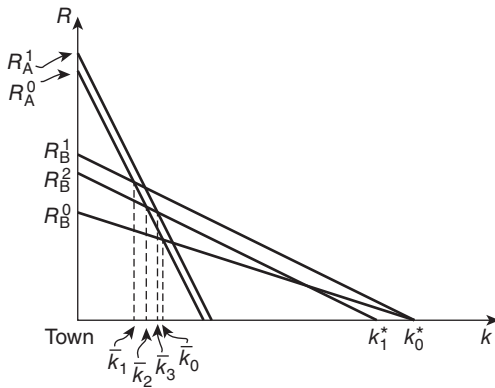


Figure 11.9 Land use with two possible crops.

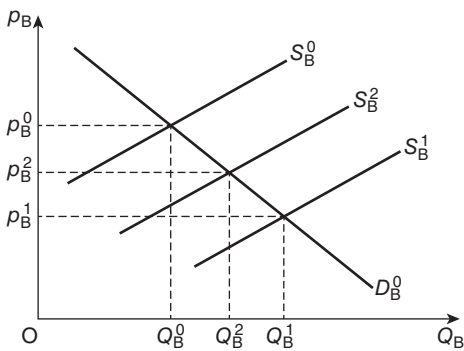


Figure 11.10 Product market responses to crop yield increase.

yield increase in crop B, possibly from the introduction (and instantaneous adoption throughout the region!) of a higher yielding variety of seed. Figure 11.9 shows the series of shifts in the land rent curves that move the region from the old equilibrium at $\bar{k}_0$ to the final, new equilibrium at $\bar{k}_3$. Figures 11.10 and 11.11 show the underlying supply-demand equilibrium changes. (I've eliminated the non-land input cost lines from Figure 11.9 for simplification.) The initial equilibrium of land uses in the region sees cultivation switch from crop A to crop B at $\bar{k}_0$ and the most distant unit of land used in growing crop B at k_0^* . The increase in Y_B (the yield of crop B) is represented by the pivot of crop B's rent curve

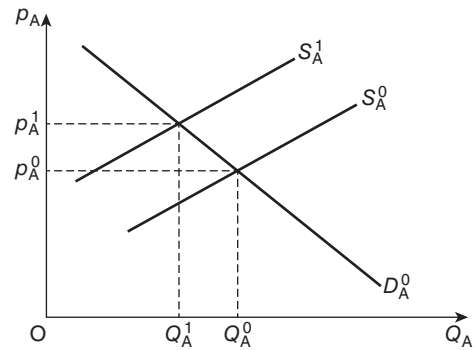


Figure 11.11 Cross-crop price effects.

in Figure 11.9 from R_B^0 to R_B^1 , keeping the initial edge of land use constant at k_0^* . This shift of the rent curve corresponds to the outward shift in the supply curve of crop B in Figure 11.10 from S_B^0 to S_B^1 , along that crop's unchanged demand curve, D_B^0 (to simplify, we assume that the income effects on both demands are inconsequential). As the supply curve slides to the right and down the unchanged demand curve, the town price of crop B falls from p_B^0 to p_B^1 , and total production of crop B increases from Q_B^0 to Q_B^1 . Notice the implication of the continuously declining town price of crop B in Figure 11.10: the separate shifts of the rent curve of crop B in Figure 11.9 – first the pivot around k_0^* , from R_B^0 to R_B^1 , then the parallel downward shift to R_B^2 – are really a single, smooth movement combining simultaneous changes in slope and vertical intercept, as the rent that crop B can offer rises at locations closer to town and falls at the most distant original areas of cultivation. Consequently, the intermediate land-use switch point in Figure 11.9, $\bar{k}_1$ will never be observed; it is just a reference point in this diagram. The initial set of new land uses that would satisfy equilibrium in the crop-B market is represented by switch point k_2 , where the new rent curve for crop B, R_B^2 , intersects the initial rent curve for crop A, R_A^0 , which is equivalent to the supply-demand intersection in Figure 11.10 that yields (p_B^1, Q_B^1) . However, the shift of the land between $\bar{k}_0$ and $\bar{k}_2$ amounts to a leftward shift of the supply curve of crop A, since some of both of

its factors of production are removed. This reduction in supply of crop A results in a higher price for that crop, with a constant demand curve for it, as shown in Figure 11.11. In the rent diagram, the town price increase results in a slight upward shift in the rent curve for crop A, from R_A^0 to R_A^1 . This shift in the rent curve, in turn, lets crop A recapture some of the land it lost initially to the increased profitability of cultivating crop B. Crop A is able to reclaim the ring of land between $\bar{k}_2$ and $\bar{k}_3$. We could pursue continuing adjustments of land allocations between crops A and B, but the movements we've shown so far suffice to show what will happen; any subsequent adjustments would result in a final switch point between $\bar{k}_2$ and $\bar{k}_3$. As we've ruled out any income effects and demand for neither crop has shifted for any other reason, we know that the new equilibrium supply of crop B will be greater than it was initially while that of crop A will be smaller: because of the change in the relative cost of producing the two crops, people will substitute their consumption a bit away from A and into B. Consequently, with the fixed-coefficients assumption about production, we also know that less land will be devoted to producing crop A, although we can't say with the information we've given ourselves whether the land devoted to crop B will rise or fall: although crop B can extend its reach closer to town than it could before the yield increase, cultivators of crop B also find it economic to cease cultivation closer to town than before too, because of the additional quantities to haul to town from the greater distances.

All our exercises with the Thünen model so far have used the fixed-coefficients assumption regarding production. That's been a handy expositional device, but some substitutability between inputs as their relative costs change is more likely. Many of the predictions of land-use changes and output responses to parameter changes carry over from the fixed-coefficients case to that of variable coefficients but some interesting differences in the spatial organization of production emerge nevertheless. The most obvious factor-price change as we move from the town toward the edge of land use is the increase in the wage-rental ratio. With

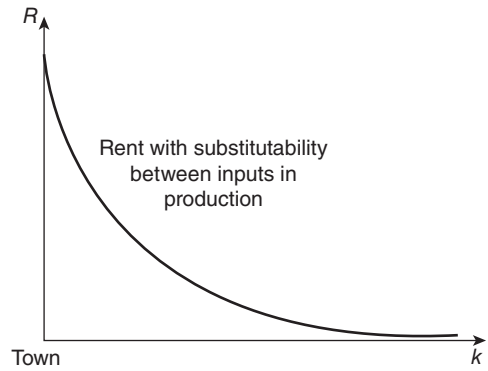


Figure 11.12 Effect of substitutability among production inputs on land rents.

a nonzero elasticity of substitution in production, cultivators will substitute land for labor farther from town, which produces a pattern of yields decreasing away from town as well. With substitutability between land and other inputs, the rent curve never actually reaches zero, so there would never be a true edge of cultivation – unless we're willing to impose something like a clearance cost on the spatially marginal unit of land, which is a reasonable thing to do; that will give a definite edge of land use. Figure 11.12 shows a rent curve for a single crop with some degree of substitutability in production. The initially rapid decline in rent reflects the rapid substitution away from land and into other inputs as the town is approached. Consequently, the shape of the rent curve is paralleled by corresponding patterns of declining intensity of non-land inputs and of outputs as we move away from the town.

So far, we've supposed that only farm outputs move around the landscape and incur transportation costs. Sometimes farms use purchased inputs that are produced in the town – or even imported from elsewhere – so some input costs may have a transportation cost gradient that increases away from the central town.

11.3.2 The bid-rent function

We can view the rent function we have been using, $R = (p_Q - tk - a)Y$, and the associated

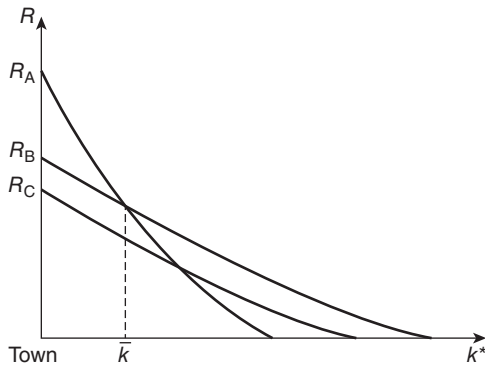


Figure 11.13 Bid-rent functions.

rent curves in the diagrams, as the maximum amount a potential user of land at location k would be willing to offer for land in the activity he is contemplating. Remember that each potential activity has a rent function associated with it, even if that activity is not chosen anywhere in the region. Thus, in Figure 11.13, crop A can outbid crops B and C for inner-zone land, crop B can outbid crops A and C for outer zone land, and crop C is not competitive anywhere in this region, given the crop's production characteristics, its local price (the demand for it locally), and the soil conditions. Nevertheless, what growing crop C *could* offer cultivators if they were to grow it is a real opportunity, just one that unsubsidized producers would not choose.

Notice that this approach to land rent has people *maximizing* rent. At first this may sound counterintuitive: why would anybody in his right mind want to *maximize* the rent he pays? That's just the catch: right now we're looking at rent from the perspective of the person who gets to keep it. Rent and profit are similar (although as we've discussed earlier, not the same), and rent maximization is conceptually similar to profit maximization. We are determining the rental price of land at the same time we are allocating it among competing uses, not just seeing how land will be allocated among users, given that it is already priced.

Let's look more closely at the derivation of the bid-rent curve. A firm's or farm's bid-rent

function is the objective function of its profit-maximization problem. Consider a profit-maximizing farm (or a utility-maximizing family farm that must account for the value of its leisure; the profit-maximizing production unit is easier, and the results are essentially the same). It produces its output, Q , with land, L , and labor, N , and it's located at distance k from the town. Its objective function is the following: $\text{Max}_{\{L,N,k\}} \pi = p_Q f[L(k), N] - wN - R(k)L(k) - tkQ$. The subscripts after the "Max" indicate that the farm has three "instruments" whose magnitudes it can adjust to bring its profit to a maximum: the quantity of land used, the labor employed, and the location. Competition will drive profit to zero. Now, rearrange what's left of the objective function to put rent per unit of land on the left-hand side: $R(k) = p_Q f(1, N/L(k)) - wN/L(k) - tkQ/L(k)$, which is exactly equivalent to the simple, fixed-coefficients form we used to introduce the Thünen model. This expression gives the maximum amount of rent per acre the farm can offer at location k while still maximizing its profit.

Suppose we've found bid-rent functions for every possible user of land for the entire region. This collection of users involves substitutions among techniques, even if factor substitution within techniques is impossible; it also will include substitutions among crops and other products and services (think of a cemetery as providing services, paid for either in cash or in foregone production). The overall envelope of maximum bids for land will appear as given to each bidder, although he or she is part of the aggregate demand for land that generates that envelope. Figure 11.14 shows such an envelope, labeled R^e , and Figure 11.15 shows a set of bid-rent curves of an individual farm, labeled R^b , each corresponding to a level of profits. The level of profits associated with the bid-rent curve decreases as we move outward in the figure: higher profits can be obtained if you can spend less for land. Like indifference curves, bid-rent curves for a given firm (or individual) and technique will not cross. Figure 11.16 shows the tangency of one of the bid-rent curves with the regional envelope of land rentals. The slope of

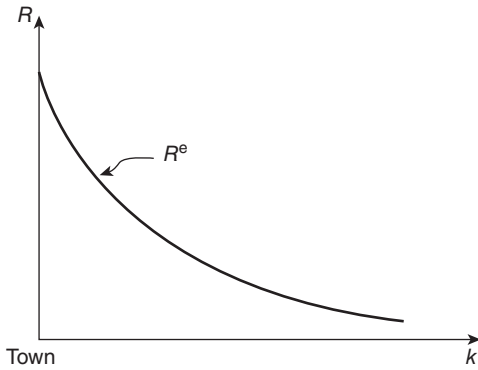


Figure 11.14 Outer envelope of bid-rent functions.

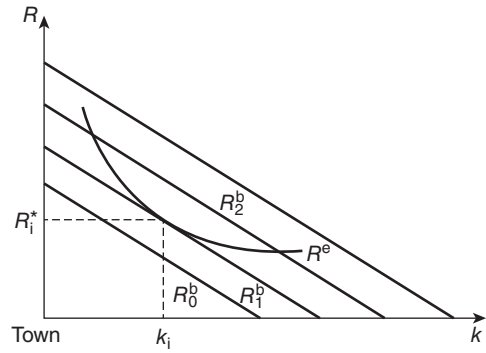


Figure 11.16 Equilibrium assignment of a firm to a location.

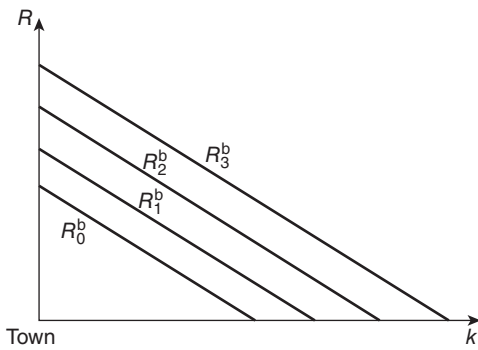


Figure 11.15 Family of bid-rent curves for a given firm and production technique.

the tangency indicates the tradeoff between rent and distance at that particular location k . The rate at which the producing firm (farm) can trade off location and land rent in its profit-maximizing production is the same at which it can make that tradeoff on the ground in the region. Farm i locates at distance k_i from town, where it pays rent (or gets to keep that rent if it owns the land) R_1^* . To earn greater profits, following bid-rent curve R_0^b , it would not be able to obtain any land. At bid-rent level R_2^b , it would lose money (or “money”).

Now, we show a short-cut technique involving the use of the bid-rent function. One of the common uses of bid-rent functions is to determine

analytically which activity will locate closer to the market (or city, or otherwise the attraction site that governs the initial point for transportation costs on local outputs). We don’t have to plot out two entire bid-rent functions to see where they will cross and which one offers higher rent on either side of the switch-point. Think about using them this way: at the switch point, $\bar{k}$, $R_A = R_B$, so $(p_A - t_A \bar{k} - a_A)Y_A = (p_B - t_B \bar{k} - a_B)Y_B$. What we want to know is which one is steeper at the location at which the rent levels are the same. To find out which rent function is steeper just from their formulaic expressions, we try adjusting distance from town, k , just a bit to obtain $\Delta R_A / \Delta k$, or the amount by which rent in crop A changes when we move a little distance right around k . Looking at the left-hand side of the last expression, the only term that will change as we move distance by the amount Δk is $-t_A Y_A \bar{k}$, which will change in magnitude by the amount $-t_A Y_A \Delta k$. Consequently the change in crop A’s rent per unit change in k , $\Delta R_A / \Delta k$, is $-t_A Y_A \Delta k / \Delta k$; so $\Delta R_A / \Delta k = -t_A Y_A$, and similarly for the change in rent to crop B at $\bar{k}$: $\Delta R_B / \Delta k = -t_B Y_B$. With due regard to the negative signs of these slopes, the question boils down to whether $-\Delta R_A / \Delta k > -\Delta R_B / \Delta k$ or $-\Delta R_B / \Delta k > -\Delta R_A / \Delta k$. This in turn hinges in this simple case on whether $t_A Y_A \geq t_B Y_B$.

Many location questions can be addressed by examining the relative slopes of bid-rent functions at land-use switch points. The alternative

bid-rent functions could refer to different crops; agriculture and various nonagricultural activities; different technologies used to produce the same crop; different contract forms for land use, possibly joined with technology and crop choices; people with different capital and skill endowments; different family structures; different preferences.

The dichotomous character of the assignment of different activity or agent characteristics to different zones, by distance from town or other attraction point, is in addition to the continuous variability of many characteristics of organization over distance that can occur within any one bid-rent curve. This may sound a bit abstract. Any single bid-rent curve characterizes the maximum rent that can be obtained under the conditions of production, preferences, and organization characterized by the bid-rent function. As long as the choices that producers can make to maximize land rent involve substitutability of various items in response to the changing relative prices or costs attributable to transportation costs, some characteristics of agricultural or other activity will vary systematically over distance on the same bid-rent curve, without any necessary reference to soil or topographic differences among locations. Of course, soil and topography may change, but their influences on choices would be separate from those of pure distance.

11.3.3 *Equilibrium in a region*

The analysis so far has been very much partial-equilibrium. Some readers may have wondered what would happen if you tried to squeeze a larger number of people into a region such as we've described. In the introduction we suggested that population growth would tend to raise demand for the locally produced crops, raising p_Q , and hence land rent, R . But think a step ahead and you'll notice that adding people, while it very well may raise p_Q and R , will have those very effects precisely because it changes the ratio of labor to land (N/L) in the region, which in turn changes production costs. The change in N/L will produce a change in the opposite

direction in w/R , and we need a way of thinking about how low w can go.

Let's introduce another analytical construct to address this issue: the "open" versus the "closed" region. We emphasize that these are constructs, and that any real region is going to fall somewhere along a continuum in its degree of "openness" used in this sense. By a "closed" region, we mean one across whose boundaries labor does not move in response to the level of the real wage (or, equivalently, utility). Consequently, the real wage, or utility, is endogenous; that is, it is determined by the interaction within the region itself of the quantities of labor and land. The population of a closed region is not endogenous, at least in the sense that it depends immediately on the level of wellbeing available in the region (although it may be endogenous to other processes outside the scope of the regional model). Defining a region as closed does not mean that the distance of the regional boundary from the center (equally a construct, but with reasonable counterparts from the observational world) is fixed. We would use the concept of the open region in cases where people, at the margin, are able to move (migrate) into and out of the region whenever the real wage or utility would tend to fall below some given level if more people were to occupy the region. That is, interregional population (labor) mobility puts a floor on the level of utility that people have to accept in any region. The population of an open region is endogenously determined, while the real wage is exogenous – "given" to the region in question by the alternative opportunities in surrounding regions. Again, the territorial size of an open region (its k^*) is endogenous, but determined in conjunction with the exogenous utility level rather than in conjunction with an endogenous utility level.

Some regions may, of course, be fixed in size, as many islands are, for all practical purposes. If an island were small enough for us to consider it a single region economically, we would have a case of a region with fixed k^* .² Whether we consider it to have been an open or closed region in the population-utility sense would remain an open question. In either case, rent at the economic edge of the region need not go to zero, but could

be bid substantially above that level. Of course, the idea of the focal point of locational attraction on an island being dead in its center, while not necessarily contradicted by the usefulness of seaports for external communication, at least needs supplementation. Some town toward the center of such an island could indeed be the primary point of locational attraction but the advantage of being located near the island's seaport could confer advantages on some producers, so we would find one gradient of rents falling away from the town at the center of the island and another, possibly overlapping rent gradient falling away from the seaport, such as Figure 11.17 illustrates. (Wieand 1987 develops such a model of a secondary urban subcenter.) The edges of the island are identified as k^* for the south side of the island and k_A^* for the north side where the seaport is located. The land rent gradient around the central town is not quite symmetric – maybe it's mountainous on the south side, increasing transportation costs and reducing rents more sharply along R_S^T than on the north side along rent line R_N^T . The seaport has no rent gradient to its north of course, but its rent gradient to the south, R_S^S , has about the same slope as the north gradient of the town, reflecting similar transportation costs in the two directions through this region, but the absolute locational advantages of the seaport are smaller than those of the town, reflected by its smaller maximum rent. Location $\bar{k}$ in the zone between the town

and the seaport marks what in all probability would be some kind of rough balance point of the orientation of activities toward the seaport to the north and the town to the south. As the extension of R_N^T all the way to the north coast of the island indicates, were it not for the seaport, producers could “serve the town's market” as far away as the north coast.

11.3.4 *Modifying the social context*

Using a centrally located market town to which all output is transported for sale may strike some readers as inapplicable to ancient Mediterranean-basin situations as the transportation surface assumption. Recall, however, that the transportation surface assumption is indeed just that: an assumption. When appropriate we can modify it. The equally heuristic assumptions that all output is taken off a farm, and that where it is taken is, of all places, a market, are equally assumptions – assumptions intended to offer the simplest presentation of the basic resource-allocation mechanisms that operate when people conduct activities across distances. They can be modified too, and we'll show some directions in which they could be modified that could be more in keeping with at least some social settings of antiquity in the Mediterranean region.

First consider a farmer who lives some distance from his fields. (For simplicity, assume the transportation surface.) Every day he must travel from his home to his fields to work.³ One of his allocation problems to solve is how far he will walk; another is, supposing that he has fields at different locations, to which fields will he devote more effort, bringing forth higher yields? One of the determinants of how far he is willing to walk is the valuation he places on his leisure, as we saw in Chapter 10. The transportation time comes at least partially out of leisure. If his production function permits him to substitute labor for land, he will tend the nearer land more intensely so he doesn't have to carry as much crop from the more distant fields.⁴

If the farmer consumes all of his own output, the transportation costs will emerge in the form of time costs and extra calories for himself and

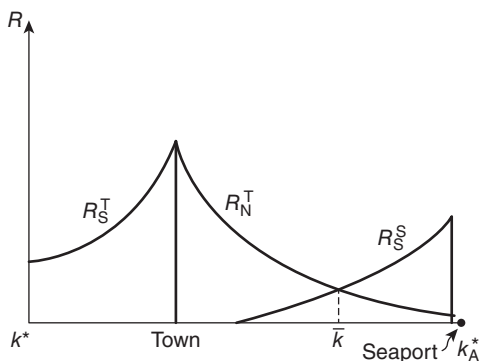


Figure 11.17 Multiple focal points for land values in a region.

any animals he uses in hauling harvested crops home. No market would be involved in either sale of output or purchase of somebody else's, but the farm household still would incur on-farm transportation costs in these two forms. Of course, with no off-farm sales, he can make no off-farm purchases; the extent to which the latter coincides with available archaeological evidence is an empirical matter.

Next, explore further the idea we introduced in the previous subsection about the central town and the seaport on the island. This example opens the subject of the endogeneity of the centers of locational attraction that we've called "the town" or "the market place." We've certainly been arbitrary in assigning the site of a town, which we've considered to be attractive enough for farmers in a region to send all of their crop outputs for sale in its markets. Why would such a center emerge in the first place? In fact, the seaport, which we considered to be subsidiary to the central town in the island example above, actually makes more sense as a site of locational attraction. As the reader should expect by now, we plead expositional simplicity and separable problems. In the following chapter, on cities, we will focus more directly on the forces contributing to the location of cities at some places rather than others. As a preview of some material from that chapter, *within* any given city, considered spatially itself rather than as a point in a region, to which our "town" in the Thünen model has been relegated, the town center (the "central business district" or CBD), or site of peak land rent, can be located arbitrarily as we've done for the entire town in the regional model or it can be determined endogenously as the envelope of many different activity patterns' locational advantages. Nevertheless, what is usefully considered endogenous at the intracity and regional levels may differ. Special geographical advantages that give additional land rents in changing transportation modes (for example, land transportation by wagon or donkey train versus ship transportation by sea or barge transportation on a river) can be powerful incentives in the competition among locations for the population agglomerations we call towns and cities.

11.4 The Location of Production Facilities

The location problem addressed in this section is where to locate a production or assembly site when inputs come from various locations, outputs are sold at other locations, and extensive use of land is not a prominent feature of the production processes involved. We will relax this last condition, but starting with it simplifies the exposition. The first subsection deals with the location of individual production facilities (we call them alternatively plants and factories, but with the same meaning). The second subsection introduces issues in the location of an industry – that is, of a group of plants that produce the same or closely substitutable goods: how will they locate relative to one another?

11.4.1 Individual facilities⁵

Once again start with the transport plain assumption, but with the qualification that some locations on the plain are the only sources of certain inputs used in the production of goods, the locations of whose demands are also specific points. In fact, we can divide inputs into two types: those whose locations are at fixed points, with zero availability elsewhere, and those that are available at equal cost everywhere – called ubiquitous resources. We need not search for a classification of things that are truly fixed-location resources and those that are ubiquitous because for some purposes we might want to call the same input fixed-location in one analysis and ubiquitous in another. Such a pair of contrasting cases might be water in, alternatively, the cases of beer brewing and grain milling. The distinction should become clear in the following discussion.

A further matter about the use of plant location theory in studying real-world observations may help the reader. This body of theory is useful for clarifying the influences of factors the analyst specifies on the location of particular production activities. When an observation appears, however, that looks anomalous from the prediction of Weberian location theory, the reader has two principal options: (i) decide the theory is wrong

(at least in the case under study) and discard it or (ii) try to figure out what factors were omitted from the analysis, or otherwise misspecified, which turned the actual result around from the predicted one. Since the first option effectively leaves us with no remaining explanatory device, we prefer the second and recommend it to the reader.

The simplest case involves a fixed-coefficients, constant-returns-to-scale, production activity that uses one fixed-location resource and a ubiquitous resource and has a single market location for its output. The unit transportation cost on the fixed-location resource is t_r and that on the output is t_Q . The plant can sell as much output as it wants at the market site at a constant price p_Q . This is a one-dimensional problem in the sense that it can be depicted as locations along a single line. Under these conditions, the factory will be located at either the raw material site or at the product market, but at no intermediate site. Trans-shipment costs are not required to produce this result, which depends entirely on the relative transport costs on the raw material and the finished good and the weight-gaining or -losing property of the final product. If the unit transportation cost on the raw material equals or exceeds that on the final good, and the production process is weight-gaining, the production will be located at the demand site to avoid the additional transportation costs on the output. If $t_r \leq t_Q$ and the production process is weight-losing, production will be located unambiguously at the raw material site. The intermediate cases ($t_r \geq t_Q$ with weight-gaining production and $t_Q \geq t_r$ with weight-losing production) require calculation of the transportation costs at the alternative sites; the site with the lower transportation cost is the cost-minimizing location. Under the assumptions about substitutability in production and infinitely elastic demand at the given product price, cost-minimization is equivalent to profit maximization.

The next case involves a production process that requires two fixed-location inputs, still in fixed proportions to each other and to output. This case could represent a metals manufacturing process, with ore and fuel the two raw materials, neither

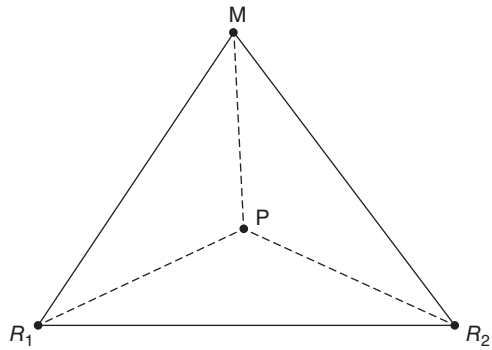


Figure 11.18 Location of a production facility between locations of market and resource sites.

being found at the market. We retain all the other conditions of the first problem. The feasible set of locations for the production facility forms a triangle, with the vertices at the two raw material sites and the demand site, as in Figure 11.18. The plant location, designated P in that diagram, is the site that minimizes the sum of transportation costs on both inputs and the output. As drawn, P is pulled off the center of the triangle, somewhat closer toward the two raw material sites and, between them, a bit toward R_2 . Changes in the raw material prices will not affect the optimal plant location; it would have the sole effect of reducing the volume of output of the plant. The two materials must be used in fixed proportion and there are no scale effects that would affect location under the present conditions. Similarly, a change in the market price of the finished product would alter the volume of output of the plant but would not give any incentive to locate closer to the market.⁶

Substitutability in production between the raw materials changes the locational invariance of the plant, as does elasticity of demand for the product at the market. Adopting either or both of these assumptions converts the simple, cost-minimization problem into a more intricate, profit-maximization problem. The choice of plant location must be made jointly with the capital decision about size of plant (which translates into volume of output per period): that is, location and production are joint decisions. A higher

transportation rate or mine-mouth price of an input will cause a prospective plant locator to plan on producing with a lower ratio of that input to the other, reducing the pull of the input whose price is higher. Think carefully about the type of price variations we're considering here – they're not transient variations around some “normal” price, but rather the long-term, expected price of a resource, and the price difference we're considering is a conjecture of a higher long-run expected price versus a lower one.

Suppose labor is a ubiquitous input, but that skills useful in a particular production process are not ubiquitous among the labor force although unskilled or less-skilled labor can be substituted for the skilled labor. The productivity of skilled labor can reduce production cost. If the optimal plant location in the two-input case, abstracting from any differential labor skills, were at site P in Figure 11.19, we can draw around P a line indicating the transportation-cost equivalent of the productivity differential that could be conferred by the skilled labor.⁷ Shifting the plant site to any location within this ring would reduce cost or raise profit. This technique can be used to study the locational effect of differential quality in otherwise ubiquitous inputs.

In some manufacturing processes, scrap can substitute for – in fact, may be preferred to – raw materials that may need more processing. Scrap typically is available at the site of markets for the material in question. Consequently the locational influence of recyclable scrap

would be to increase the locational pull of the demand site.

Return to our earlier example of water as a fixed-location resource in one case and a ubiquitous resource in another. Although the Mediterranean basin in general might not be an ideal area in which to think of water as a ubiquitous resource, settlement sites invariably found water for drinking, if not in forms and quantities for navigation. While water used in brewing may make unique contributions to the flavor of the beer, at least some quality of beer consumed in a population will be able to be profitably produced with whatever quality of local water is available. In this sense, when considering the location of brewing sites as closer to grain growing areas or to principal consuming sites such as cities, the city is likely to win out as the predominant brewing location.⁸ Water flowing at sufficient velocity to turn a grist mill will be found in far fewer locations than water of the characteristics we have proposed for run-of-the-mill beer. For purposes of thinking where such mills would be located, water is usefully thought of as a fixed-location resource, although there may be several fixed locations for potential mill locators to consider.

There is a taxonomy of the location factors important to a line of production that provides a useful check-list for thinking about specific cases. A production activity, or an industry, is said to have a resource orientation, or to be resource oriented, if such raw materials form a large cost share and the transportation costs on them are substantially higher than on the output. A market-oriented industry may have low cost-shares of raw materials, or use ubiquitous raw materials, and have high transportation costs on finished products. Production processes that can use scrap as a substitute for raw materials will have their market orientation strengthened. A production process in which labor, particularly skilled labor, is important is called labor-oriented; it is likely to have its optimal location pulled away from both markets and raw material sources and toward locations where the skilled labor force lives, possibly for historical reasons that have nothing to do with recent economic conditions.

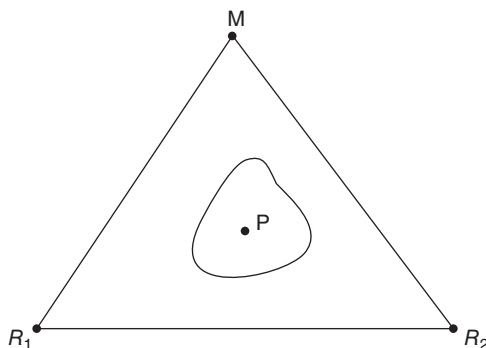


Figure 11.19 Productivity differentials around a production facility site.

11.4.2 Industries

Where will an entire industry of production facilities locate? This question is too general to answer, but we can break it down and get some meaningful answers to partial questions. First, let's restate that the product produced by this industry is identical across plants: their outputs are perfect substitutes for one another. Second, the plants are owned and operated by perfectly competitive agents; that is, they are neither a monopolist's multiple outlets nor a small number of facilities of recognizable competitors such as would characterize duopoly or oligopoly. Now, we could ask where such an entire set of plants would locate if none of them were already sited, or we could pose the more practical question of what their sales would be from their current locations. We are inclined toward the latter question, but before addressing it, we need to characterize the spatial distributions of both consumers and producers.

We have a market area as our subject if producers are concentrated, in either a single point or multiple points, and consumers are more or less evenly distributed over space. The locational question becomes how the consumers are allocated among the sellers. If, in contrast, consumers are concentrated at a single point, possibly in a city, and sellers are dispersed over some territory, we have a matter of the supply area of the consumers: from which producers will they buy? This case, if you think about it for a moment, is essentially that of the Thünen model.

In the market area problem, since the products of the different producers are identical, each consumer will be content to purchase from the seller offering her the lowest price. The price facing each consumer will depend on the pricing policy of the producers. Let's suppose that each producer offers his product at his factory gate at its marginal cost of production and consumers travel to his factory to purchase the good. Consumers all face the same transportation costs in all directions, so they will travel to the producer for whom the consumer's combination of factory price and own transportation cost is lowest. Note that this allows producers with different marginal costs to coexist within a region, as long

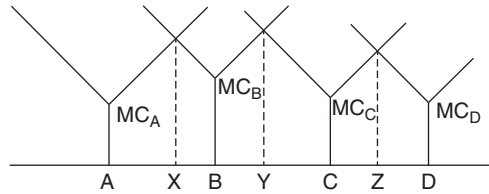


Figure 11.20 Production sites, product delivery costs, and market areas.

as distances and transportation costs are great enough to shelter the higher-cost producers from differences in consumers' transportation costs. This is an immediate and important departure from the aspatial, purely competitive model. Figure 11.20 shows four producers' locations on the horizontal axis and prices on the vertical axis. Producers A, C, and D all have identical marginal costs, but producer B has a marginal cost half again as large. The oblique lines originating from the marginal cost points are lines of delivered prices – consumers at each location on the horizontal axis “deliver” one unit of the good each period to their own home sites by going to the producer's location and getting it. Producer B still captures a market area in the zone from X to Y although its factory price is higher than the other producers' prices, but as you can see, the half of that market area from B to Y is less than half the distance between the two producers' sites.

Figure 11.21 shows a plan view of the market areas of three producers. The circles around each of the P_i producer locations indicate the farthest distance any consumer would travel for the good these producers sell. The overlaps of the circles indicate zones of consumers that either seller could service fully in the absence of the other seller, but consumers will travel to (purchase from) the seller closer to them, as indicated by the straight lines bisecting the zones of overlap. This diagram may strike some readers as looking like the beginnings of central place theory, with its overlapping market areas chopped off around the edges to yield hexagons. This is indeed the market area component of central place theory, but this treatment alone omits important allocations the consumer must make in his purchase

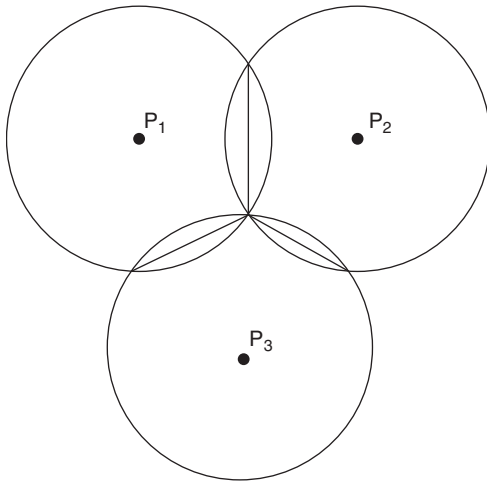


Figure 11.21 Least-cost distances from market centers and overlapping potential market areas.

behavior as well as the economies of scope and scale in consumer travel.

Buried within Figure 11.20 is the assumption that either the production levels of each producer are fixed or that they produce under constant costs. To see this more clearly, think of the supply curve of each of the producers in that diagram, as we show in Figure 11.22(a). Each increment

in the quantity supplied raises the marginal cost of all units produced. Each quantity Q of output can be associated with the quantity of this good that consumers within the firm's market area are willing to purchase at a delivered price. As the producer serves a larger market area, the delivered price to more distant consumers has two sources of increase: a higher marginal cost, faced by all closer consumers as well, and the transportation cost. With any elasticity at all in the demand for this product, the ability to sell to more distant consumers is coming at the expense of some foregone sales to nearer consumers, and the producer could devise a pricing policy other than f.o.b.-mill pricing (the consumer picks up the product at the mill gate and pays marginal cost) that would let him maximize his profit.

Next, we can approach an answer to the question of where producers would locate if they were not already in production. The plants depicted in Figures 11.20 and 11.21 could be squeezed together until the quantities they produced were at the minimum points of their average cost curves, as shown in Figure 11.23. The output level q^* in Figure 11.23 corresponds to the demand in the smallest market area viable from the producer's perspective. Even in circumstances in which a large number of producers are already

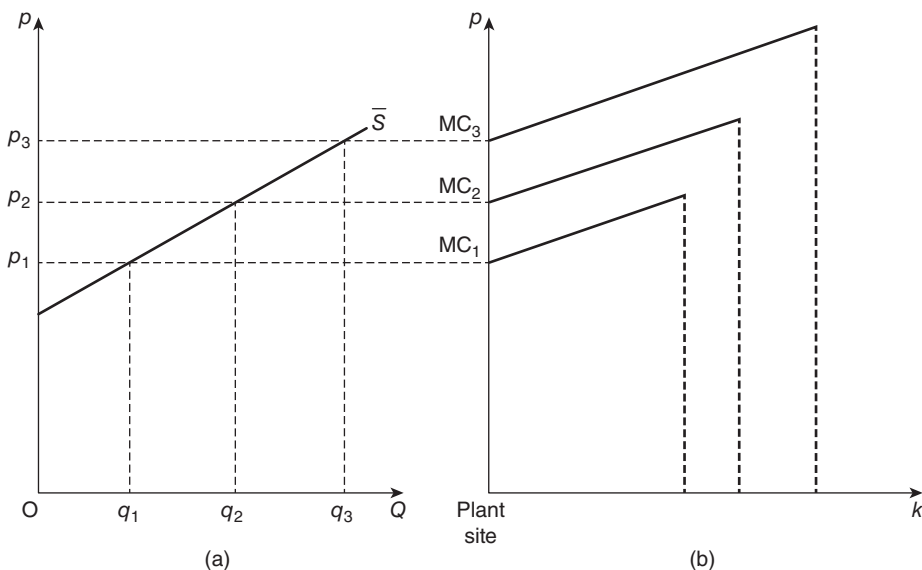


Figure 11.22 (a) Supply curve for a good supplied over a spatial market. (b) Sources of increase in delivered cost as output and market area increase.

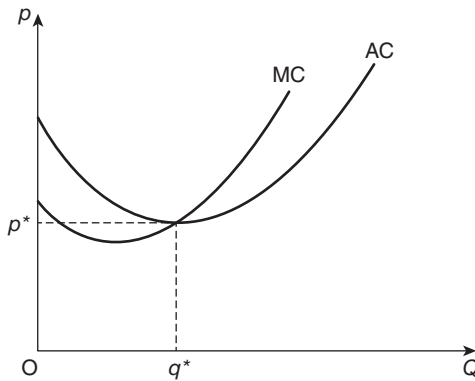


Figure 11.23 Determination of smallest viable market area.

located, as long as new producers can enter the market-territory, this condition characterizes just how many more producers would be able to enter.

11.5 Consumption and the Location of Marketing

The third major strand of location theory commonly is referred to as central place theory, but the element of central place theory that has received the most attention – the hierarchy of markets – is actually the most weakly developed of the components in the location theory of markets and consumption. Accordingly, we develop the individual components of this theory and leave the hierarchy bits for a short discussion toward the end of the section. Spatial consumption theory adds many elements of realism to the theory of demand as we've treated it so far. The household production model forms the implicit basis of much of the spatial extension of demand theory: in addition to the price of a product consumers face when they arrive at a place where the product is sold, they have to get there themselves and get the product back home, which are costly activities. If the product is delivered to them at their own doorstep, they still have to pay for the delivery. Either way, multiple goods could be purchased on the same trip, reducing the transportation cost assignable to any one of the goods. While such bundling makes it more difficult to say exactly how much any one of the goods cost in total, the bundling yields some

insights about demand interactions between goods that don't have any direct relationships to each another, which in turn yield predictions about the structure of marketing efforts.

We open the section with an expanded conception of transportation costs, then turn to the important concept of the frequency of purchases. With these concepts, we turn to the resource allocation efforts on the parts of consumers, then sellers. We close the section with some general comments on hierarchies of market places.

11.5.1 *The structure of transportation costs*

So far we have specified transportation costs as strictly proportional to distance traveled and quantity of material carried. While that's a reasonable first approximation, transportation costs in shopping trips are more complicated. First, depending on the mode used – foot, animal, vehicle – the travel cost may not be particularly sensitive to the number of units of a good carried. For example, a cart may have capacity to carry more than a single item at the same cost in time, animal feed, and wear and tear on the cart. Even with bulk material, the cost of various quantities in the same wagon may increase in less than full proportion to the weight carried. These are economies of scale. Second, even if a single unit of a good is purchased, a bag swung over the shoulder may be able to carry single units of several different goods at pretty much the same cost as that of a single good. This number of different goods jointly purchased, or services jointly consumed, on a single trip gives rise to economies of scope. Altogether, the transportation cost on a shopping trip may be a function only of the distance traveled, and not of the quantity of goods carried from market to home. This trip cost would vary by the mode of travel, which choice itself might be influenced by the quantity or number of items the consumer intends to purchase. The contemporary automobile or light truck has accentuated these characteristics of shopping transportation, but even the transportation methods of antiquity would have been able to generate these economies, if at a lesser scale.

The effect of such a transportation cost structure is to make the purchase of multiple goods cheaper

than the purchase of the same goods separately. It will affect where consumers shop, where sellers locate and what they offer, sellers' pricing policies, and the dispersion across sales outlets of prices for identical goods.

11.5.2 The shopping tradeoff: frequency versus storage⁹

The consumer's shopping problem is to find the optimal bundling of goods and services and determine the optimal frequency of shopping. The two are related. If storage were costless, consumers would make purchases less frequently than they do. As the case is, they can save on transportation expenses by buying in bulk and storing, but the storage is costly too. There is a minimum-cost combination of shopping frequency and quantity of goods stored that minimizes the sum of transportation and storage costs. Storage costs are affected by the price of the goods stored (the capital cost – interest cost – of inventory), real depreciation (deterioration) of the commodity, and the cost of constructing and maintaining appropriate storage space. There will be tradeoffs also between the deterioration permitted in stored goods and the quality of storage space constructed.

We work through a simple example of how this tradeoff works. Let s_i be the cost of storing one unit of good i per time period. Suppose for the moment that there are no joint purchases. The price of a unit of the good at the sales site is p_i , the quantity of it purchased and stored is x_i , the frequency of purchase is ϕ_i , and the round-trip travel cost to acquire the good is t . Then the transportation and storage costs together in any time period are $\phi_i t + s_i x_i / 2\phi_i$; the first term is the travel cost per trip times the number of trips made, and the second term is the storage cost of quantity x_i for half of the average holding period (assuming the quantity x_i is consumed at an even pace, beginning with the full amount at the beginning of the period defined by ϕ_i and drawn down to zero at the end of the period, leaving an average of half of the initial quantity stored for the entire period). Notice that as ϕ_i gets larger, the first term gets larger and the second

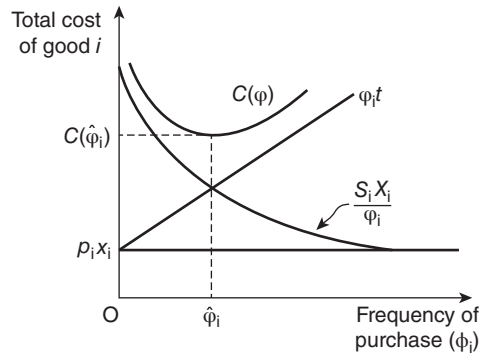


Figure 11.24 Tradeoff between frequency of purchase and total product cost. Adapted from Stahl 1987, 797, Figure 6. Reprinted with permission of Elsevier/North-Holland.

term gets smaller. Consequently, we can manipulate the value of ϕ_i to find the frequency that minimized this total travel-plus-holding cost. Doing that we get a cost-minimizing frequency of $\phi_i = (s_i x_i / 2t)^{1/2}$. This optimum shopping frequency is larger with higher storage costs and larger quantities consumed and is smaller with larger transportation costs.

Now we consider the full cost of getting good x_i by adding its purchase price to the transportation and storage costs. We use a function notation for this cost, $C(x_i, \phi_i, p_i, t)$, to indicate that the cost of getting amount x of good i is a function of the quantity obtained, the frequency of purchase, the market price of the good, and the travel cost to get it. Then $C(x_i, \phi_i, p_i, t) = p_i x_i + \phi_i t + s_i x_i / 2\phi_i$. Figure 11.24 draws the components of this cost versus the frequency of purchase. The direct purchase cost, $p_i x_i$, doesn't change as purchase frequency changes, so it's a straight line. The storage cost falls asymptotically (that is, it gets smaller and smaller but never crosses the axis that it's approaching) as ϕ_i increases, and the travel cost rises linearly with frequency. Add the three costs together and the total acquisition and holding cost is u-shaped, meaning that it has a minimum at some particular shopping frequency, which we showed in the previous paragraph.

As we have already found the cost-minimizing shopping frequency in terms of storage cost, quantity purchased, and transportation cost, we can substitute that expression for the optimal frequency into the original cost function and eliminate shopping frequency as a separate variable in the cost function for good i : $\hat{C}_i(x_i, p_i, t) = p_i x_i + (2s_i x_i t)^{1/2}$. Examination of the right-hand side of this expression tells us that, although increases in both purchase quantity, x_i , and transportation cost, t , as well as the holding cost, s_i , increase the cost of acquiring and holding this good long enough to consume it, they increase it at a decreasing rate.

Now, using the same cost-function framework, let's turn to the issue of economies of scope. Suppose that we have two market sites, A and B. Market A offers only good i , while market B offers both goods i and j . The consumer can reduce her consumption cost by purchasing both goods at B as long as the cost of buying the two goods jointly at B is less than the cost of buying good i at market A and good j at market B. Using our notation developed just above for minimized cost (minimized by optimizing the shopping frequency), we set up the question as $\hat{C}_i(x_i, p_i^A, t^A) + \hat{C}_j(x_j, p_j^B, t^B) > \hat{C}_{i+j}(x_i, x_j, p_i^B, p_j^B, t^B)$; the first term on the left-hand side is the lowest cost of buying good i from market A, the second term on the left-hand side is the lowest cost of buying only good j at market B, and the right-hand side is the lowest cost of buying both goods at B. What conditions will make the specified inequality true? Suppose that the transportation costs to A and B are the same ($t^A = t^B$) and that the market price of good i is the same at both places ($p_i^A = p_i^B = p_i$). The expression for the optimal-frequency storage costs when the goods are bought separately at the two markets is $\sum_{i,j} [(s_k x_k)^{1/2}]$ (where subscript "k" can take on the values i and j for the two goods), and when they are bought jointly is $(\sum_{i,j} s_k x_k)^{1/2}$. The sum of the square roots will be larger than the square root of the sum, so the differential storage costs will provide the cost saving of joint purchases necessary to induce the consumer to shop at market B for both goods. This consumer will purchase good i at market A

only if the price of good i is much lower at A than at B, or if the travel cost to A is substantially lower than the cost to B, *and* if she demands good i in large quantities. Even in that case, she may still buy some of her good i at market B.

Economies of scale for consumers in hauling larger quantities of any given commodity cause them to be attracted to market sites that offer discounts. They can divide travel costs over a larger quantity of the good. Correspondingly, demand in an entire market (group of consumers) will increase more than proportionately to a decrease in the price of any single good. Economies of scope in transportation produce interdependencies in market demands – that is, the demands of groups of people – for different goods that do not exist at the level of the individual consumer's preferences. The market demand for one good at a particular location depends not only on its own price but on the prices of other goods available there – not simply because of the pure substitution effect in each consumer's demand system but because of the income effects of avoided transportation costs. A real cost-of-living decrease derives from an increased choice of commodities at a market site.

Following the concept of compensating differentials introduced in earlier chapters, prices can be higher at sites offering greater choice, and in fact at least some of them will have to be to equilibrate shopping between these sites and those offering narrower arrays of goods. However, the entire array of prices is a package deal of sorts, and if even one price at a site gets too far out of line, a consumer (indeed any number of consumers) may cease shopping at that market entirely because such a price change affects her demand for all other goods at that site, whether they are individually substitutes for or complements to the good whose price has changed.

Barry Lentnek, Mitchell Harwitz, and Subhash Narula (Lentnek *et al.* 1981; Narula *et al.* 1983; and Harwitz *et al.* 1983), in a series of papers in the early 1980s, have considerably fleshed out the consumer's side of the central place problem – frequency of shopping for various goods at various distances, consumers' time

and out-of-pocket costs, the cost of keeping inventories of goods at home once they've been purchased, the ability to combine transportation costs of multiple purchases. These models assume that the locations of markets offering the different goods are known and fixed, so they are short-run, partial-equilibrium analyses. The timescale involved in the consumer choices is much higher-frequency than the investor choices involved in establishing and changing market sites, which makes for quite an intricate general equilibrium problem.

11.5.3 *Aggregate demand in a spatial market*

In the introduction to economies of scale and scope in consumer transportation, the focus was on the individual consumer. The existence of transportation costs can convert goods that are substitutes at the level of the individual consumer's preference function into complements at the aggregate level. In this subsection we show how this occurs and discuss its implications. For this demonstration, we'll literally add up the demands for some specific good by consumers who are spread over some market area and see how price variations affect the total quantity demanded of the good. Assume that consumers travel to the market place to purchase the good. We use an indirect utility function to help express the maximum distance that any consumer will travel. To refresh your memory, the indirect utility function, v , is a function of goods prices and income: $v(\mathbf{p}, Y^*)$, where the bold $\mathbf{p}$ represents a vector of prices – that is, an array of prices for a variety of goods – and Y^* is income net of shopping travel costs: $Y^* = Y - tz$, where z is distance from the market to a consumer's location. As prices increase, the value of v decreases, and as net income increases, indirect utility rises. Now, we define the maximum distance that consumers are willing to travel to purchase this array of goods as $\bar{z}(\mathbf{p}, \bar{v})$, where $\bar{v}$ is the minimum level of utility the consumer has to accept. Each individual's demand function is $\mathbf{x}(\mathbf{p}, Y^*)$, where the bold $\mathbf{x}$ indicates that this is a compact representation of the consumer's system of demands for an entire array of goods x_i . Next, we can give ourselves a special form of demand function that lets us separate the effect of prices and income

on demand into net income at any location z times a demand function based only on price: $\mathbf{x}(\mathbf{p}, Y^*) = Y^* \mathbf{x}(\mathbf{p}, 1)$. This will let us aggregate demand in an entire market by multiplying the demand function independently of location times the net income a consumer at location z has, and adding up these demands over all locations z from the site of the market to the farthest distance any consumer will travel to get any of this array of goods, $\bar{z}$: $\mathbf{X}(\mathbf{p}, \bar{v}) = 2\mathbf{x}(\mathbf{p}, 1) \sum_{z=0}^{\bar{z}(\mathbf{p}, \bar{v})} Y^*$ (note, of course, that Y^* changes with location z , and we are effectively adding demands over different locations from $z = 0$ to $z = \bar{z}$; the 2 gets in here because we're adding line areas on either side of a market).

When the price of one good changes (call it good i) at this market, the demand for any other good (say, good j) has two components to its subsequent change: $\Delta X_j / \Delta p_i = 2\bar{Y}\bar{z}\Delta x_j / \Delta p_i + 2\bar{Y}\mathbf{x}(\mathbf{p}, 1)\Delta\bar{z} / \Delta p_i$.¹⁰ This expression says that the change in the total quantity of good j demanded by consumers distributed over this market area is composed of two effects, the first one essentially a substitution effect between goods i and j at the level of the individual consumer, and the second one a change in the radius of the market area over which consumers are willing to travel once the price of one of these goods changes. In the first term $\bar{Y}$ is a variant of the location-specific income, Y^* ; we're just multiplying that spatially adjusted net income at each location, times the radius of the market area (times 2), times the individual's substitution effect. The second term multiplies the demand at the edge of the market area, $\bar{Y}$, times the change in the market area radius caused by the change in the market price. Remember that the market price is one of the determinants of $\bar{z}$, the maximum travel distance in the market area; when any p_i changes, $\bar{z}$ changes in the opposite direction. For a positive change in any price the second term is always negative, but the first term is positive or negative, depending on whether goods i and j are substitutes or complements. The implication of these two effects acting at the aggregate level is that the market relationship between two goods can be complementary even if the goods are substitutes at the level of the individual consumer if the substitution effect is smaller in absolute value than the market area effect. In such a case, a decrease in the price of some good i would

not only increase the demand for that good, but it would increase the demand for its substitute good, j , by enlarging the area over which j is purchased from this market. Additionally, the offering of some other good (call it good k) at this market place would increase the demand for substitutes for that good as well as for complements to it because the number of customers buying at this market would increase.

From the seller's perspective, consumers' economies of scope in transportation let sellers confer positive externalities on one another by locating nearby or changing their prices, or any other actions they might take. An individual seller who sells an array of products could internalize these effects. In fact, while it is correct to think of a monopolist selling at a higher markup than a perfectly competitive seller, a monopolist selling multiple goods may sell at a lower price the same goods that competitive sellers offer as single-product sellers. The monopolist will always sell complementary goods for less than competitive, single-good sellers would ask for the same goods individually. By internalizing the market-area effect, a monopolist also can offer lower prices than perfect competitors could on goods that are substitutes, as long as the goods aren't too close substitutes. While each single-good competitive seller can correctly assess the substitution and market-area effects of its own price changes on its own demand, they don't (can't; need not) account for the negative effect of the market area change on another competitive seller's sales.¹¹

11.5.4 *Hierarchies of marketplaces: central place theory*¹²

Central place theory began with the observation that there existed a series of markets of different sizes offering larger and smaller arrays of products and services. The importance of the transportation cost savings available from multiple-good purchasing and the value of comparison shopping were attributed considerable importance but the behavioral theory underlying the full model of the hierarchy of retail sites has never been worked out satisfactorily (it's quite difficult). On the sellers' side of the behavior is the choice of location relative to other sellers – it's reasonable to assume evenly distributed consumers for this part of

the problem. Sellers compete with one another for customers, and locating so as to be closer to more of the consumers than a competitor is an important element in retail locational strategy. However, as we've noted, these sellers, even direct competitors, also benefit from one another's presence; sellers of different goods benefit even more unambiguously from locating near one another.

On the buyers' side, considering all the purchases to be made, there will be a distribution of single-purpose and multi-purpose shopping trips: sometimes you need only a carton of milk and some bread; other times you need to get a carpet cleaned and some new shoes as well. A sufficient number of buyers' single-purpose trips (or more generally, trips on which a restricted array of goods is to be purchased) will keep a small ("higher order") market (central place) in business. The optimal frequency of purchases of relatively low-price, regularly consumed, rapidly depreciable (for example, milk spoils) goods will be higher than the optimal frequency for goods with higher prices (for example, consumer durables) that deteriorate less quickly. Consequently the frequency of purchase is linked accordingly with the order of a central place, higher order places being associated with higher shopping frequencies.

Returning to the sellers' perspective, markets need to be able to serve a minimum number of consumers to break even; this number of consumers translates into a minimum-size market area (called the "minimum range" or threshold). The "maximum range" is the greatest distance a consumer will travel to purchase a particular good, but as we've noticed above, this distance depends on the array of goods sold along with any particular good. The maximum range of a "higher order" central place could be greater than the maximum range of the individual good with the largest market area if the positive externalities among sellers let them offer these goods at lower prices than if they were sold separately. On the other hand, the buyers will be willing to pay a premium for at least some of these goods because of the cost savings they experience in transportation and from the real benefit of comparison shopping.

Sellers will adjust their locations and their scales of operation, along with their pricing strategies (percentage mark-up over cost) as part

of their profit-maximization strategies. Whether the traditional, zero-profit equilibrium condition will emerge from this spatial problem is not fully understood at present, but there is at least a tendency to squeeze out “excess profits.” In a final, spatial equilibrium (the hypothetical long run that may never actually be reached but towards which forces of scarcity tend to push agents, just as in the aspatial treatment of these problems), the landscape will be fully served with overlapping market areas of different-sized central places, all with hexagonal market areas, which are just sets of overlapping circles with the overlaps cut off by straight lines. The market areas of some higher order central places will cross the market areas of the next lower order of central places, which seems to be a result of more interest for its geometry than its behavioral significance. Christaller noticed empirically three orders of progression of central places by size in the region he studied, which he called the marketing principle (“ $K = 3$ ”), the transportation principle (“ $K = 4$ ”), and the administrative principle (“ $K = 7$ ”). The “ K number” indicates the number of central places of order $n + 1$ associated with each central place of order n ; thus a system of central places developed under the marketing principle would have, say, one central place of the highest order (where brain surgery can be bought or rented), three central places of the next lower (second) order, nine of the third order, and so on. To count the lower order places associated with any higher order place, you have to count the segments of the lower order places that overlap the market area of each corresponding higher order place. For example, with the $K = 3$ system, there is one second-order market area of the first-order central place that is contained fully within the first-order place’s market area, and there are second-order places at each of the six vertices of the hexagonal first-order market area. These six second-order places and their market areas have to be divided among the three higher order places into whose market areas they fall (see Figure 11.25). The six second-order places are each divided among the three first-order places into whose market areas their own market areas fall, giving two second-order places associated with each first-order place; then we add

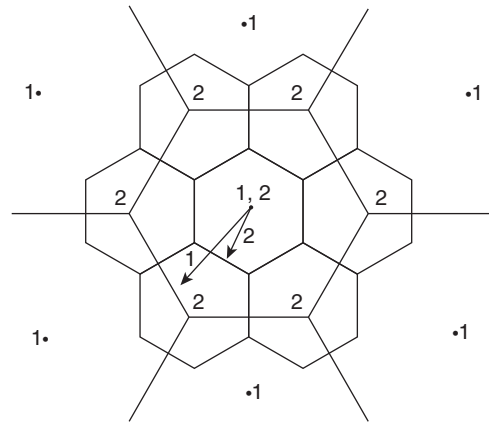


Figure 11.25 Central place theory: different orders of market places and market areas.

the second-order place that lies wholly within the first-order place’s market area, to get a total of three second-order places associated with each first-order place (result: $K = 3$). Lösch derived theoretically many more possible rates of increase of central places than the three Christaller identified empirically in southern Germany.¹³ The behavioral foundation of these “systems” also has not been satisfactorily learned yet. Consequently, inferences about what was important at a time and place (region) on the basis of the progression of market towns (central places) by size are difficult to justify with any confidence.

11.5.5 Periodic markets

Central place-theoretic concepts have been applied to marketing systems composed of locations that are not open for business on a daily basis but rather convene at regular intervals for given durations, known as periodic marketing systems. The basic force driving these less-than-full-time markets is a deficiency of demand within a given area, for whatever reason, to support full-time markets. De Ligt (1993) has brought an array of these periodic markets from the Roman Late Republic Period to the Early Imperial Period to the attention of ancient historians, as one aspect of the urbanization of Italy during those periods.¹⁴ Most of the empirical literature on

periodic marketing derives from Africa and Central America, and De Ligt's pushing back the dates of observations, however rough, by some 2000 years certainly adds to the empirical base of that literature. The central place concepts offer a reasonable accounting for these markets and their gradual disappearance, or rather merger into permanent facilities, with urbanization.

11.6 Transportation

It is customary to think in terms of transportation modes, which are combinations of mobile equipment and the fixed infrastructure on which or with which the "rolling stock" is operated. The transportation modes of the ancient Mediterranean basin would have been water, vehicle, animal, and the human foot. Within each of these modes, at virtually any place and time, alternative forms of transportation existed within each of these modes: different types and sizes of ships and boats, different types of land vehicles, different animals, even different types of human foot transportation (for example, porters and self-transportation). The modes would have tended to differ in terms of their ratios of fixed to variable costs, their absolute levels of fixed costs, and their speeds. Agents with different personal and business characteristics would have used the modes in different frequencies.

Rather than deal with the ancient transportation facilities by mode, which is rather more suitable for specialist works in remains, textual descriptions and artistic representations,¹⁵ which focus on technology, we divide transportation into what we consider the economically relevant components of infrastructure, equipment, and pricing. Infrastructure tends to be public goods while much of the equipment tends to be privately owned and operated. So far in our analysis of locational issues, we have withheld transportation prices from actual allocational forces: we have assumed them to be given and exogenous to the demands placed on the transportation systems, when in reality, the cost – and accordingly the price – of transportation will respond to the same forces of demand and supply that cause

the quantities and prices of food and clothing to change.

11.6.1 Infrastructure

Much transportation infrastructure tends to have public-good characteristics: excluding users is difficult, particularly with roads and city streets; one consumer's use of it subtracts little if anything from the availability for other consumers, although traffic congestion is characteristic of an impure public good; large scale tends to send consumers who could use only a small fraction of the minimum size facility looking for others to share the cost, which characterizes harbor facilities. It's easier to restrict consumers' access to bridges but the conveyance over obstacles that bridges offer frequently strike populaces as services that ought to be available to all, bringing in the merit aspect of public goods.

Consequently, transportation infrastructure tends to be paid for, if not necessarily built by, whatever passes for the public sector in a society, be it the royal treasury or the public assembly. Whether they are actually built by public agencies, such as a royal engineering corps, or private contractors on the public payroll is more an accident of time and place than necessity. Paying for infrastructure falls into two components: building it in the first place and maintaining it after it suffers the wear and tear of users. Unless users can be made to pay their marginal valuation for these transportation services, they will tend to overuse these infrastructure facilities, equating the marginal benefit of their privately provided transportation inputs to the average cost of using them (essentially the "fisheries problem" in which freely available, unpriced fish in the ocean are depleted by the excessive use of privately provided ships to catch them).¹⁶

Unless tolls or something along that line can be extracted from users, the maintenance on infrastructure that provides its reliability will have to be financed through taxation. Poorly maintained transportation infrastructure lowers the productivity of private transportation, as users have to devote extra resources to compensations of various sorts, ranging from allowing extra travel time as a buffer, to greater wear and tear

on vehicles, to extra port time for ships loading and unloading. Will the bridge carry the load on these carts? Will the road ahead be washed out after the last rain, or will previous maintenance be likely to have provided for timely diversion of runoff? While it is common to hear of roadway toll systems decried as unwarranted extractions of surplus from innocent traders, such tolls have the advantage of letting the road's users pay for keeping highwaymen at bay and the roads themselves in usable condition, both of which cost real resources. To the extent that transported goods cost more to final consumers as a consequence, they are simply paying part of the transportation cost. The alternative financing would reduce the incomes of some or all consumers by the amount required to finance the roads, regardless how much road-transported consumption they enjoy.

The infrastructure for water transportation would have been harbor facilities such as wharves and docks, any dredged harbor areas, possibly smoothed-off beach areas in locations and times where ships were beached rather than docked.¹⁷ Warehouses near sea and river ports may be considered as part of water transportation infrastructure, whether they are privately or publicly owned. Ports compete with one another through the efficiency of their facilities and the prices they charge for their use. Ports also have hinterlands for material they ship out as well as goods they take in. A reduction in port charges will expand the size of the hinterland, at constant overland transportation costs.

A port authority has to decide how much it will charge shippers for use of its facilities. There's little reason not to believe that port authorities would maximize their profits if they knew how, so let's assume they do (and did). They will face an elasticity of demand for their facilities in general. The smaller the throughput through a port, the more elastic is the demand, and the greater the throughput the less elastic. A port facility may face shippers with considerable monopoly power – for example, the Prince of Tyre at the height of the Egyptian New Kingdom may have serviced an Egyptian shipping fleet that carried not only its own nationals' cargoes but a large proportion of other nationals' cargoes in the

region as well and hence could have set shipping rates to maximize their revenue (Katzenstein 1973, 47). By the eleventh century B.C.E., the winds of power had shifted in the region, as the Tale of Wenamun illustrates, and the Prince of that time evidently had greater pricing power (Lichtheim 2006, 224–230). In either type of case, the port price that the Prince of Tyre chose would have influenced the shipping rates charged by the Egyptian fleet. In general, a profit-maximizing port authority would charge a higher facility price to competitive shippers than to shippers with monopsony power. However, if the port authority has some interest in maximizing the welfare of its own nationals who ship goods out of the port, the port fees should be lower than the price that would maximize only the port authority's profits. This is because the monopsony shipper's freight rates respond quite sensitively to the port charges, so a lower port charge lowers ocean freight charges by even more, and local producers farther from the port can afford to send exports through it while the producers who already could afford to ship goods through the port benefit from the higher farmgate prices they receive.¹⁸

Roads, city streets, and bridges (as well as toll stations¹⁹ where such might have existed) are the infrastructure of land transportation. It would be common for many city streets to have been developed originally by nearby private building owners, but the benefits most such streets would confer on others with little or no interest in those buildings would relatively quickly make them effectively public goods: "Hey, let's take this short-cut!" In light of the importance of animals in land transportation, it might make sense to include some public wells in transportation infrastructure, just as service stations are included in contemporary inventories of highway transportation infrastructure.

Pricing is less of an issue with most of these components of land transportation infrastructure than was the case with ports because exclusion is easier in ports. This does not mean that financing is a less important matter, for it certainly is not, but general taxation, and possibly *corvée* labor, would be the more common mechanism. In another

sense financing the land infrastructure would have been more problematic because there would have been less feedback from user-imposed costs such as wear and tear to continued use unless poor maintenance and deterioration were allowed to be the (inefficient) messenger. It is also possible that only the roads and bridges that were immediately profitable to the government coffers would have been maintained, or even built: indirectness of the links between general improvements in wellbeing and government revenues could have discouraged such government expenditures; whether governments in the Mediterranean basin in the second millennium B.C.E. were able to make this connection is an empirical question, possibly with textual answers. Of course, small rural towns may have been able to identify their residents' collective valuations of local roads and assemble the resources from those residents to construct and maintain local road infrastructure that would have been beyond the means of central governments to evaluate.

11.6.2 *Equipment*

Transportation equipment is likely to have had a considerably higher proportion of private suppliers than did infrastructure, although the more expensive end of the equipment – particularly ships – may have been owned by governments relatively more often than less expensive equipment simply because of their scale relative to private incomes. The use of all types of transportation equipment, from ships to shoes, is more easily excludable than is infrastructure.

Government agencies, individuals, and groups of individuals all may have owned various types of transportation equipment at various times and places. Ships, being the largest and most expensive of the mobile equipment, would have had the most restricted ownership but even private individuals are known to have owned entire ships by the Classical Period in Greece (Demosthenes 32.2, “Against Zenothemis,” in Murray 1936, 175, 179; Demosthenes 33.1–12, “Against Apaturius,” in Murray 1936, 203–209; Demosthenes 34, “Against Phormio,” in Murray 1936, 233, 243; Demosthenes 45.64, “Against

Stephanus I,” in Murray 1939, 223; Demosthenes 56, “Against Dionysodorus,” in Murray 1939, 191–192, 195–197; Isager and Hansen 1975, 65–66, 73–74; Reed 2003, 98–132). The potential for economies of scale in ships could have given governments advantages in owning (or financing the construction of) the larger ships in the ancient inventories.

As in the contemporary world, military requirements appear to have formed most important sources of demand for technological changes, the most obvious being in ships and chariots; demands for better carts for the quartermasters may simply be less observable because of the lack of distinction between civilian and military supply carts. And who can say about footgear? Would actively serving infantrymen have been more likely to wear boots, sandals or other foot protection than people of similar income levels in civilian occupations? However, we should not overlook the civilian demand for technological change in cargo ships coming from merchants who stood to gain from economies of scale in shipping. Undoubtedly, the physical risks to cargoes at sea would have dampened the employment of technical capabilities to make larger ships, but recalling Hesiod's recommendation to farmers to put their cargoes on larger ships, the proportion of larger to smaller ships may have been elevated by the relatively greater risks of smaller ships foundering. Arnaud (2011, 73) notes several advantages of smaller ships (20 to 50 metric tons capacity): their access to a larger number of harbors and the greater ease of finding enough shipments to fill them. Whether the greater construction and operational expenses of larger ships was an additional reason to prefer smaller ships depends on the possible economies or diseconomies of scale in construction (including financing) and operation.

11.6.3 *Pricing of transportation services*

In Thünen's original model of the determination of agricultural land rent, he posited a horse pulling a cart full of grain from a farm gate to the central town, with the horse eating some of

the grain on the way. The image is an excellent reminder that transportation costs real resources that could be put to other uses. There is a marginal cost of transportation just as there are marginal costs of shoes and wheat.

The demand for transportation is a derived demand, just as is the demand for labor. Accordingly it is amenable to study with the Marshall–Hicks laws of derived demand. If we have grounds for believing that transportation costs formed a larger proportion of delivered prices of many goods in the ancient Mediterranean basin, we should expect there to have been a more elastic demand for transportation than is typically the case in the contemporary industrial world.

Most of our locational analysis has used very simple forms of transportation rates: prices linear in both quantity carried and distance hauled. The example of modifying this linearity assumption in the analysis of retail consumption behavior and location should alert the reader to the potential power of nonlinearities in transportation costs to affect behavior and locational patterns, the latter of which are capital (investment) decisions. Ocean shipping typically has freight rates that decline with distance carried. Rate structures may look like step functions instead of linear functions, as shown in Figure 11.26. The cost per ton may be unchanged over some considerable zone. Step-function freight rates are more likely to be quoted prices rather than costs incurred; they will reflect the cost of more finely grained pricing within the

transportation industry but whatever their cause, they will have real allocative effects on shippers facing them.

Fragility aside, bulkier, low-weight goods (Corinthian aryballoi) typically incur higher ton-mile rates than do heavier, more compact goods (grain and olive oil). Competition plays a role in transportation pricing: freight rates on routes with fewer carriers will be higher than where competition among carriers is stronger.

In noncompetitive situations, shippers may be able to transfer transportation costs from one group of consumers (with more elastic demands for the delivered good) to another (with less elastic demands). Alternatively, a producer paying for the delivery of his own goods may find it profit-maximizing to charge less than the sum of the mill-price and the full transportation cost, a practice called freight absorption. Another practice which has given economists interested in industrial competition a workout for some 50 years is called basing-point pricing. In such a delivered pricing system (consumers pay to have the product delivered to them – they don't go pick it up themselves), the delivered price quoted a buyer at, say, location A by a seller at, say, location 1, is the mill price of the seller at location 1 plus the transportation cost on the shipment from some other site (the basing point) – say location 3 – to demand site A. This transportation cost may be more or less than the real transportation cost from supply site 1 to demand site A. The variance in the delivered price from the real c.i.f. cost ("cost-insurance-freight") may protect some less efficient producers, possibly using older production facilities. The practice will tend to reinforce regional patterns of supply-demand interaction, encouraging demanders at some locations to patronize sellers at particular locations that otherwise would not be least-cost to them. At the very least, the sellers in an industry practicing basing-point pricing have to agree on the pricing system and the basing points; this has been interpreted as *prima facie* evidence of collusion of one sort or another. Nevertheless, the practice has been defended as offering some seasonal price stability when the seasonal pattern of orders cannot be filled by customary, least-cost

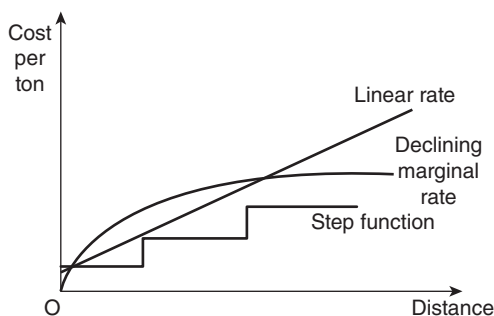


Figure 11.26 Types of transportation rates.

suppliers (Haddock 1982). Pricing systems such as these certainly have the ability to affect the economic development of regions – the regional distribution of industries and the levels of real incomes in different regions.

Transportation is a particularly capital-intensive activity. An increase in the demand for transportation services that raised its price (as contrasted with leaving freight rates unchanged) would raise the real rental on capital by a larger proportion than the transportation rate increase. It would also depress the wage-rental ratio by more than the proportional change in transportation prices relative to other prices in the economy. This result is independent of the relative magnitude of the transportation sector in the economy.²⁰

11.7 Suggestions for Using the Material of this Chapter

The intensity of use of land depends on both large-scale and small-scale locations relative to some especially desirable sites. For example, consider a city a hundred miles from Rome and another city the same size 50 miles from Rome, and sites within both cities. If we think of land use gradients within both cities, the one 50 miles from Rome can be expected to have a higher peak density than the one 100 miles away, and the density gradient in the more distant city can be expected to have a flatter slope.

Historians and archaeologists have discovered the Thünen model. Some have actually used it as a guide in research (for example, Morley 1996, 58–68 and subsequent applications). Don't be put off by the unrealistic assumptions of the model – the uniform plain, transportation surface, and so forth. Those assumptions just clear the thickets to facilitate understanding of the basic forces at work. The model still gives pretty good empirical predictions. In application don't hesitate to allow for interruptions in theoretically smooth gradients of density, price, or whatever, when terrain features intervene.

Typical production facilities unearthed by archaeologists include metallurgical workshops

and pottery kilns. Now and then a purple dying facility may be identified by a cluster of murex shells near a seashore. Sites of grape crushing for wine production and olive pressing have been somewhat less common. The latter three types of facility are cases of perishable raw inputs needing immediate treatment before they spoil. Those facilities can be expected to be located quite close to the source of the raw material. The locations of metallurgical facilities are sometimes more puzzling. Scrap probably was not a typical major source of input in the earliest antiquity, although as time progressed, it surely could have been. Scrap would have been more common around larger urban centers, where the scrapped goods had been used. The other major input for metallurgy would have been fuel. If fuel and metal source (say, ore in the absence of ample scrap) had been located close together, metallurgical production sites could be expected to be located nearby. Transportation of bulky ore, which would have been reduced greatly in weight and increased greatly in value per pound, would have told ancient metallurgists to minimize hauling it around. That said, hauling by sea would have been considerably less costly than hauling by land. Fuel – exclusively wood at the time – would have been as bulky as ore to transport but lighter, suggesting that it could have moved further, other factors being the same. Carrying finished metal products to final markets would have been less costly in terms of transportation resources than transportation of the raw materials close to the final markets would have been. Ingots could be transported to locations where they were hammered into final products. The bronze and tin ingots found in various shipwrecks around the eastern Mediterranean and the Aegean suggest this pattern of producing final metal products close to their markets was a common practice. If finishing metal products required close consultation with final consumers, the importance of location of the manufacture of those implements close to final markets would have been reinforced. Against this background, Betancourt's excavation of the Early Minoan bronze working facility at Chrysokamino on eastern Crete is a mystery (Betancourt 2006). The ore was brought

into Crete, probably from Cyprus, smelted on site, and the bronze then manufactured into final consumer goods at that relatively remote site. Other ancient metallurgical sites should be examined with these locational principles held

in mind. Clearly the Chrysokamino location was profitable for a long time; the reasons behind the success of its location, and those of other similar smelting sites, warrant examination.

References

- Adams, Robert McC. 1981. *Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates*. Chicago IL: University of Chicago Press.
- Arnaud, Pascal. 2011. "Ancient Sailing-Routes and Trade Patterns: The Impact of Human Factors." In *Maritime Archaeology and Ancient Trade in the Mediterranean*, edited by Damian Robinson and Andrew Wilson. Oxford: Oxford Centre for Maritime Archaeology, Institute of Archaeology, pp. 61–80.
- Bennathan, Esra, and A.A. Walters. 1979. *Port Pricing and Investment Policy for Developing Countries*. New York: Oxford University Press.
- Betancourt, Philip P. 2006. *The Chrysokamino Metallurgy Workshop and Its Territory*, Hesperia Supplement 36. Princeton NJ: The American School of Classical Studies at Athens.
- Bromley, R.J., Richard Symanski, and Charles M. Good. 1975. "The Rationale of Periodic Markets." *Annals of the Association of American Geographers* 65: 530–537.
- Chisholm, Michael. 1962. *Rural Settlement and Land Use; An Essay in Location*. New York: John Wiley & Sons, Inc.
- Christaller, Walter. 1933. *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen*. Jena: G. Fischer.
- Christaller, Walter. 1966. *Central Places in Southern Germany*. Translated by Carlisle W. Baskin. Englewood Cliffs NJ: Prentice-Hall.
- Crouwel, J.H. 1981. *Chariots and Other Means of Land Transport in Bronze Age Greece*, Allard Pierson Series, Vol. 3. Amsterdam: Allard Pierson Museum.
- Crouwel, J.H. 1992. *Chariots and Other Wheeled Vehicles in Iron Age Greece*, Allard Pierson Series, Vol. 9. Amsterdam: Allard Pierson Museum.
- DeGraeve, M.-C. 1981. *The Ships of the Ancient Near East (c. 2000–500 B.C.)*, Orientalia Lovaniensia Analecta 7. Leuven: Departement Oriëntalistiek.
- De Ligt, L. 1993. *Fairs and Markets in the Roman Empire: Economic and Social Aspects of Period Trade in a Pre-Industrial Society*. Amsterdam: J.C. Gieben.
- Dunn, Edgar S., Jr., 1954. *The Location of Agricultural Production*. Gainesville FL: University of Florida Press.
- Haddock, David D. 1982. "Basing-Point Pricing: Competitive vs. Collusive Theories." *American Economic Review* 72: 289–306.
- Hall, Peter G., ed. 1966. *Von Thünen's Isolated State*. Translated by Carla M. Wartenburg. Oxford: Pergamon.
- Harwitz, Mitchell, Barry Lentnek, and Subhash C. Narula. 1983. "Do I Have to Go Shopping Again? A Theory of Choice with Movement Costs." *Journal of Urban Economics* 13: 165–180.
- Isager, Signe, and M.H. Hansen. 1975. *Aspects of Athenian Society in the Fourth Century B.C.* Odense: Odense University Press.
- Jacobsen, Thorkild, and Robert McC. Adams. 1958. "Salt and Silt in Ancient Mesopotamian Agriculture." *Science* 128: 1251–1258.
- Jones, Dilwyn. 1995. *Boats*. Austin TX: University of Texas Press.
- Jones, Donald W. 1978. "Production, Consumption, and the Allocation of Labor by a Peasant in a Periodic Marketing System." *Geographical Analysis* 10: 13–30.
- Jones, Donald W. 1991. "An Introduction to the Thünen Location and Land Use Model." In *Research in Marketing, Supplement 5. Spatial Analysis in Marketing: Theory, Methods, and Applications*, edited by Avijit Ghosh and Charles A. Ingene, 35–70. Greenwich CT: JAI Press.
- Katzenstein, H. Jacob. 1973. *The History of Tyre; From the Beginning of the Second Millennium B.C.E until the Fall of the Neo-Babylonian Empire in 538 B.C.E.* Jerusalem: Schocken Institute for Jewish Research.
- Landström, Björn. 1970. *Ships of the Pharaohs: 4000 Years of Egyptian Shipbuilding*. Garden City NY: Doubleday.
- Lentnek, Barry, Mitchell Harwitz, and Subhash C. Narula. 1981. "Studies in Choice, Constraints, and Human Spatial Behaviors." *Economic Geography* 57: 362–372.
- Lichtheim, Miriam. 2006. "The Report of Wenamun, P. Moscow 120." In *Ancient Egyptian Literature: The New Kingdom*. Berkeley CA: University of California Press.

- Littauer, Mary. 1973. *Wheeled Vehicles and Ridden Animals in the Ancient Near East*. Leiden: Brill.
- Littauer, Mary, and J.H. Crouwel. 1973. "Early Metal Models of Wagons from the Levant." *Levant* 5: 102–126.
- Lorimer, H.L. 1903. "The Country Cart of Ancient Greece." *Journal of Hellenic Studies* 24: 132–151.
- Lösch, August. 1944. *Die raumliche Ordnung der Wirtschaft; eine Untersuchung über Standort, Wirtschaftsgebiete und internationalen Handel*, 2nd edn. Jena: G. Fischer.
- Lösch, August. 1954. *The Economics of Location*. Translated by Wolfgang F. Stolper. New Haven CT: Yale University Press.
- Meiggs, Russell. 1973. *Roman Ostia*, 2nd edn. Oxford: Clarendon.
- Morrison, J.S., and R.T. Williams. 1968. *Greek Oared Ships 900–322 B.C.* Cambridge: Cambridge University Press.
- Morley, Neville. 1996. *Metropolis and Hinterland; The City of Rome and the Italian Economy 200 B.C. – A.D. 200*. Cambridge: Cambridge University Press.
- Morrow, Karen D. 1985. *Greek Footwear and the Dating of Sculpture*. Madison WI: University of Wisconsin Press.
- Murray, A.T. 1936. Translator, *Demosthenes IV. Private Orations XXVII – XL*. Loeb Classical Library. Cambridge MA: Harvard University Press.
- Murray, A.T. 1939. Translator, *Demosthenes V. Private Orations XLI – XLIX*. Loeb Classical Library. Cambridge MA: Harvard University Press.
- Narula, Subhash C., Mitchell Harwitz, and Barry Lentnek. 1983. "Where Shall We Shop Today? A Theory of Multi-Stop, Multi-Purpose Shopping Trips." *Papers of the Regional Science Association* 53: 159–173.
- Raban, Avner, ed. 1985. *Harbour Archaeology; Proceedings of the First International Workshop on Ancient Mediterranean Harbours, Caesarea Maritima, 24–28 June, 1983*, BAR International Series. Oxford: British Archaeological Reports.
- Reed, C.M. 2003. *Maritime Traders in the Ancient Greek World*. Cambridge: Cambridge University Press.
- Reisner, G.A. 1913. *Catalogue Général des Antiquités Égyptiennes du Musée du Caire, Nos. 4798–4976 et 5034–5200: Models of Ships and Boats*. Cairo: IFAO.
- Samuelson, Paul A. 1983. "Thünen at Two Hundred." *Journal of Economic Literature* 21: 1468–1488.
- Schörle, Katia. 2011. "Constructing Port Hierarchies: Harbours of the Central Tyrrhenian Coast." In *Maritime Archaeology and Ancient Trade in the Mediterranean*, edited by Damian Robinson and Andrew Wilson, 93–106. Oxford: Oxford Centre for Maritime Archaeology, Institute of Archaeology, pp. 93–106.
- Stahl, Konrad. 1987. "Theories of Urban Business Location." In *Handbook of Regional and Urban Economics*, Vol. 2. *Urban Economics* edited by Edwin S. Mills. Amsterdam: North-Holland, 790–813.
- Steffy, J. Richard. 1994. *Wooden Ship Building and the Interpretation of Shipwrecks*. College Station TX: Texas A&M University Press.
- Thünen, Johann Heinrich von. 1826. *Der isolierte Staat in Beziehung auf Landwirtschaft, und Nationalökonomie*. Hamburg: Friedrich Perthes.
- Tzedakis, Y., St. Chrysoulaki, Y. Venieri, and M. Avgouli. 1990. "Les routes minoennes: Le route de Coiromandres et la contrôle des communications." *Bulletin de correspondance hellénique* 114: 43–65.
- Tzedakis, Y., St. Chrysoulaki, S. Voutsaki, and Y. Venieri. 1989. "Les routes minoennes: Rapport préliminaire, défense de la circulation ou circulation de la défense?" *Bulletin de correspondance hellénique* 113: 43–75.
- Weber, Alfred. 1909. *Über den Standort der Industrien*. Tübingen: J.C.B. Mohr.
- Weber, Alfred. 1929. *Theory of the Location of Industries*. Translated by Carl J. Friedrich. Chicago IL: University of Chicago Press.
- Westerberg, Karin. 1983. *Cypriote Ships from the Bronze Age to c. 500 BC*. SIMA Pocketbook 22. Gothenburg: Paul Åströms Förlag.
- Wieand, Kenneth F. 1987. "An Extension of the Monocentric Urban Spatial Equilibrium Model to a Multicenter Setting: The Case of the Two-Center City." *Journal of Urban Economics* 21: 259–271.
- Woytowitsch, Eugen. 1978. *Die Wagen der Bronze- und frühen Eisenzeit in Italien*. PBF XVII.1. Munich: C.H. Beck.

Suggested Readings

- Greenhut, Melvin L. 1970. *A Theory of the Firm in Economic Space*. New York: Appleton-Century-Crofts.
- Hoover, Edgar M. 1948. *The Location of Economic Activity*. New York: McGraw-Hill.
- Hoover, Edgar M. 1971. *An Introduction to Regional Economics*. New York: Knopf.

Notes

- 1 Thünen (1826) (2nd edn, 1842, 3rd edn, 1863) (English translation: Hall (1966)). For more recent expositions see Dunn (1954); Samuelson (1983); Jones (1991).
- 2 An island such as Crete very well might be large enough that this analogy wouldn't be appropriate for it; indeed, some of its major valleys might be considered as separate open regions at some periods of its prehistory and separate closed regions at other periods – although we can't really say at present which periods are likely to have been the open-region periods and which the closed-region periods!
- 3 Of course, sometimes farmers working particularly remote sites will set up rude shelters near their dispersed land – frequently shepherds tending remote pasture – and return home only seasonally, or at least at some frequency far less than daily.
- 4 Chisholm (1962, Chapter 4) offers evidence from Europe, Africa, and Asia on typical travel times to plots, percentages of total cultivation devoted to plots at various distances from the farmstead, and yields on plots at different distances from the farm. The data come from the period prior to extensive availability of motorized transportation in these regions.
- 5 Although there were a number of predecessors, primarily in the German literature, the most lasting influence on plant location theory is Weber (1909) (English translation: Weber (1929)). Commonly referred to in contemporary writing as “the Weberian theory or model,” it serves as the point of departure for most models of the location of production (which has expanded into the study of how space and locational choices affect production behavior), and as the origin of many of the questions that contemporary models endeavor to answer.
- 6 It might strike the reader that the constant-returns-to-scale assumption makes the volume of output of the plant indeterminate but the location of a plant involves the choice of what size plant to build as well as where to build it. Once the facility size is fixed, at least one factor (the capital in the plant) is not free to vary, which will give rise to increasing cost as the rate of output is increased.
- 7 This contour is called the “critical isodapane”: isodapane because it is a contour of locations of equal unit transportation cost; critical because this particular isodapane delineates the region of profitable relocations from unprofitable ones.
- 8 Of course, there may be some small proportion of the beer consumed in a city that is brewed in the far-off mountains with its pure spring water and hauled to the city at extraordinary cost to quench the beer-thirst of those most willing to pay for the transportation cost on all that water. *Hoi polloi* are likely to get most of their beer from local brewers using the water from the local wells or canals.
- 9 The models used in this subsection and the next are taken from Stahl (1987).
- 10 The exact definition of $\tilde{Y}$ is $Y - \bar{t}z/2$.
- 11 Evidence for planned clustering of retail shops comes from Ostia, the port city of Rome, from the first century B.C.E. through the second century C.E. and later. While most shops were set in rows along the frontage of public buildings and apartment blocks, they are occasionally found grouped together in independent architectural units. Such shopping bazaars have not been located in contemporary Pompeii (Meiggs 1973, 273–274).
- 12 The two original contributions are Christaller (1933) (English translation: Christaller (1966)); and Lösch (1944) (English translation: Lösch (1954)). Christaller was a geographer; Lösch an economist.
- 13 Lösch developed his central-place model to include production as well as sales at each of the central places, although he did not try to incorporate the problem of production with fixed-location resources into his model. Christaller's model took production as given.
- 14 De Ligt (1993, pp. 4–13) cites a reasonable array of the theoretical literature on periodic markets, which appeared between the mid-1960s and the early 1980s, generally fading thereafter into exhaustion. He cites more of the economically minded literature, produced mostly by economic geographers after being prompted by Asianist anthropologists, than a rather contradictory literature from anthropologists and cultural geographers, for example, Bromley *et al.* (1975). In addition to full-time traders in periodic markets, an issue in the literature has been the participation of part-time as well as full-time traders, linking trading to other occupations, a subject on which I have made a very modest contribution myself (Jones 1978).
- 15 The works are plentiful. For example, on ships, Morrison and Williams (1968); Westerberg (1983); Reisner (1913); Landström (1970); Jones (1995); DeGraeve (1981); Steffy (1994). Harbor facilities: Raban (1985). Land transportation:

- Crouwel (1981, 1992); Woytowitsch (1978); Littauer and Crouwel (1973), 102–126; Littauer (1973); Lorimer (1903); Morrow (1985).
- 16 By “privately provided ships,” I do not intend a contrast with government-provided ships, which existed generally only as warships, not fishing vessels, but to indicate that the owners and operators of these vessels are concerned only with private, not social, costs and benefits.
 - 17 Schörle (2011, 100–102) discusses the existence and use of private harbor facilities at seaside villas, which she calls “villa harbours,” during the Roman period in both Italy and Asia Minor, one of their advantages being the ability of shippers to evade taxes at the more established, public ports.
 - 18 For analysis of a number of topics in port pricing and investment, see Bennathan and Walters (1979). The previous discussion of port pricing is adapted from Chapter 5.
 - 19 The structures found along the apparent Minoan roads in eastern Crete and interpreted as guard houses could have served as toll stations as well (or instead) (Tzedakis *et al.* 1989, 1990).
 - 20 To see this, you can think of the increase in the relative price of the small transportation sector’s output as a decrease in the relative price of the output of the other, large sector of the economy (we could call it agriculture). Agriculture being a relatively labor-intensive sector, a decrease in its price hits labor relatively strongly – that is, since much of the cost underlying the agricultural price is labor cost, a decrease in the price of agricultural output must disproportionately reduce the cost of labor; there aren’t many places for labor to go (that is, there aren’t a lot of substitution possibilities), so its wage has to drop. The magnification effect of the output price changes on the input prices doesn’t depend on the relative size of the sectors as long as we’re analyzing the economy as a whole – that is, transportation relative to everything else.

Cities

12.1 Cities and their Analysis, Modern and Ancient

Cities have been the subject of love-hate relationships for almost as long as people have been expressing opinions about them. This chapter will endeavor to sidestep these lovers' quarrels and give the reader some insights of immediate use in analyzing the resource allocation aspects of ancient cities from whatever sources remain. This section introduces the focus of the remainder of the chapter and addresses what may be a number of points of skepticism regarding the usefulness of contemporary thinking about cities and urbanism to its ancient manifestations.

12.1.1 *Classifying cities*

Both the historical study of cities and the more policy-oriented analysis of "contemporary urban problems" tend to classify. Classification in itself isn't bad, as long as the analytical purpose of the classification, if there is one, is kept in clear view, but for the twentieth century and some of the previous one the tendency has been to classify cities "as they really are," neglecting the necessarily partial comprehension of any classificatory system. Thus we have borne the terminologies

of the parasitic city for the European Middle Ages and much of antiquity, the consumer city for various parts of the same long time periods, and so forth. Of course, "parasitic" is an epithet rather than a neutral, informative adjective. A more patient assessment of the sort of situation that term seeks to name might characterize the portion of the consumption conducted in a city or group of cities that is produced in it (or them); that division of observations might lead in turn to characterizations of the location of ownership of activities relative to their actual undertaking; and before we know it we have a complexity of information that is difficult to capture in a simple, metaphorical name. Engels' proffered replacement of Pirenne's and Weber's (Max, not Alfred) concept (endorsed by Finley) of the consumer city as "the model" of the ancient city has generated a number of useful insights about ancient cities in general and Corinth in particular, and if we are to evaluate "visions" of cities, ancient or modern, it probably is an improvement over its century-older rival, but in many ways it is no more general despite its improved accuracy of observation and analysis on a number of points.¹ We can look at these taxonomies of cities as the best thinking that the late nineteenth century had to offer, but the twentieth century saw much original thinking about cities, and much

of it can be applied usefully to the analysis of ancient urbanism.²

This chapter will use classifications of cities in several instances, several different classifications in fact, according to specific analytical needs. The classifications will be avowedly partial, characterizing only one or possibly a very few dimensions of a multidimensional phenomenon. When we classify cities in section 12.5 according to what they produce, we will have excluded already 50 to 60% of what they produce before we make the classificatory assignments! When we move from the purely analytical use of classification to observational uses – and surely the two uses are closely related – the classification becomes something of a statistical representation, which we can expect to fit general trends but not all cases, generally because of the importance of omitted variables for particular observations. This may sound rather abstract right here, but when we get to the actual practice later in the chapter, you'll find yourself forearmed.

12.1.2 Characteristics of cities

In our relatively observationally based approach to the city, it's useful to start with the kinds of observations that would lead us to think we've found a city. Cities are agglomerations of people and activities. The spatial density of various indicators is a prime characteristic of urbanism. Large populations within restricted areas yield high population densities. What's "large," and what's "high"? Clearly these two descriptors are best considered relatively. Contemporary cities may have residential population densities from ten thousand to several hundred thousand per hectare. The entire populations of many ancient cities were smaller, but those sites were no less recognizable as cities (to either their contemporaries or us) than are contemporary New York and Calcutta.

The density of structures in cities parallels the density of populations. Seventy to eighty percent of the surface area of a city may be covered with structures of one sort or another, from private residences to public edifices to infrastructure such as streets, bridges, walls, aqueducts,

and so forth. Little of the sort appears in the countryside. The density both of the population and structures contains implications about activities and social mechanisms that occur in cities. These two observations, in fact, stand at the core of the analytical constructs developed to study cities.

The greater diversity of people found inside cities than outside them, and their density of living and working together, yield an intensity and frequency of interactions greater than would occur in other settings. The opportunities for people to affect one another outside of markets appear nowhere greater than in cities. Of course, thinking back to the economic characterization of people affecting one another outside of markets brings us squarely to the concept of the externality. Cities are full of externalities, and it is these externalities that make cities such frightfully productive places relative to other sites in the civilizations of their times. People meet with people of different backgrounds and who possess different information sets and take away from their interactions new ideas – ideas for which they paid only indirectly if at all.

We should include the importance of space as a consideration in most private and public decisions in cities as one of the defining characteristics of cities. True, space is important in agriculture, but without downplaying its importance there, we should emphasize the importance of space in cities. Space being both limited and delimited in cities, means of acquiring it and efficiency of using it are of first-order importance. Overcoming spatial separation inside cities is as important as using space; indeed, it requires devotion of scarce space to transportation systems to overcome spatial separation of interacting locations.

12.1.3 What goes on in cities

It will be useful to think closely about just what activities occur in cities. Three principal categories of activities are especially prominent. First, people live in cities; the remains of houses are our tangible evidence in the ancient cities. Between 30 and 50% of land in contemporary cities is devoted to residential structures, even when

construction technology permits multiple-story dwelling units such as the high-rise apartment building. Second, they consume in cities – most people find it convenient to conduct most of their consumption where they live, so even if we have little direct evidence of ancient consumption activities in cities, we can rest assured that they were there. Third, production occurs in cities and surely did in the ancient cities as well as the medieval and the contemporary ones. There is some direct evidence for production in cities, ranging from remains of foundries and forges, various craft workshops identified by tools and remains of materials, to apparent warehouses. Some readers may be inclined to distinguish between “production” and the “distribution” of which warehousing is a factor of production but we recommend thinking of that activity as productive, just as is manufacturing or agriculture.

The agglomeration of production activities in cities is *prima facie* evidence for the existence and importance of scale economies. Without scale economies of some sort, there would be no advantage to locating so many people in such close proximity to one another, especially considering the concomitant disadvantages of crowding.

Between this subsection and the last, we’ve introduced just about the full range of subject matter involved in the economy of cities. The phenomena introduced here give very substantive targets for investigation, permitting us to keep our feet on the ground, so to speak, when introducing tools for analyzing the cities of the ancient Mediterranean basin. We turn now to the joining of archaeological and textual evidence to the contemporary intellectual constructs we will introduce in the remainder of the chapter.

12.1.4 *Ancient observations and contemporary analytical emphases*

The evidence we possess on the ancient cities is considerable, despite the fact that it really is just “remains.” Archaeology has delivered us individual buildings, both residences and public

structures, cities, and entire systems of cities.³ Much reconstruction can be accomplished with the sensitive interpretation of these remains, sometimes supplemented by textual remains as well. One purpose of the following sections of this chapter is to help form new questions to put to this material. Occasionally, these ideas may offer additional tools to improve responses to questions whose answers have remained elusive. For example, estimates of populations for many of the ancient cities have been constructed as what are effectively engineering estimates: estimated number of people per “standard” dwelling times number of “standard” dwellings in some measured area, times the total area in the city (sometimes just within the walls). The scholars making these projections have been aware of their limitations: that the number of people per building is pretty much a guess, and that different buildings surely would have had different numbers of residents; that building density surely varied across the ancient cities. The urban population density gradient, which will be developed in section 12.4, uses a behavioral basis for estimating the residents per unit of land along transects across a city; with some effort and no more heroic assumptions than are involved in the engineering projections, application of this concept could offer better buttressed city population estimates.

Let’s turn to the obvious differences between the cities of today, for whose observations the ideas presented here have been developed, and the cities of the ancient Mediterranean region. First, the extent of urbanization of entire national populations is far greater today than then: 65–80% of national populations of Western Europe and North America being urbanized today, relative to a likely high of 7–10% in antiquity. Correspondingly, the share of agriculture in the output of the ancient economies was considerably higher than today: probably 80–90% in antiquity, compared with 3–6% in the industrialized European and North American nations today. Cities are suited for nonagricultural pursuits, so the scope for both their range of activities and their overall economic importance would have been far more circumscribed than the cities

of today. Third, both building and transportation technologies have enormously influenced cities around the world over the past century, permitting unparalleled expansion upward as well as outward. Fourth, the scale of productive activities in today's cities and the transportation technologies of even the nineteenth century permitted substantial spatial separation of residence and workplace. A much larger proportion of ancient city dwellers surely worked closer to home than do today's workers, including many sharing residential and work space. Nevertheless, the proportion of ancient urban workers who spent a smaller share of their total time traveling between home and work may or may not have been smaller than among contemporary urban workers.

The idea of a "housing market" may be anathema to many contemporary scholars of antiquity, and surely the concept should be approached carefully. Taking an indirect route to housing markets, consider housing implications of health in ancient cities. Until fairly late in the nineteenth century C.E., cities were pretty unhealthy places, often requiring net immigration just to maintain their populations. Granted that the great, epidemic diseases appear to have descended on the West toward the end of antiquity (surely Athens in 430/29 B.C.E. was a precursor, as was the Antonine Plague of the second half of the second century C.E. in Rome and its empire), poor sanitation and haphazard drinking water quality, not to mention their exacerbated effects in dense populations with high proportions of poverty, would have imposed a drag on natural rates of population increase in all ancient cities throughout antiquity. Scheidel (2003) paints a grim picture of living conditions and disease in Rome and Scheidel (2004, 15–17) considers quantitative effects of additional mortality from disease on immigration necessary to maintain the city's population.⁴ This is a recipe for housing turnover; surely not every new immigrant built his or her dwelling. Could we rely on "urban tribalism," whereby towns and villages maintain a constant flow of their populations to the same decimated quarters of the cities in which their cousins just died, effectively keeping the housing "in the

family?" Housing markets, whatever their institutional structures may have been, surely were needed to allocate people and accommodations to one another.⁵

12.2 Economies of Cities

In this section we look at what is peculiarly urban about activities that could be undertaken anywhere in a country but in fact occur in cities, either sometimes or always. The crucial factor in the distinctive urban contribution, we will find, is space, which we learned how to think about in economic terms in the previous chapter. The first subsection addresses urban production. The second includes both production and consumption activities as they are affected by the spatial aspect of cities. The final subsection offers the first of the partial taxonomies we promised in section 12.1.

12.2.1 *Scale economies in production*

In Chapter 1, on production, we touched on the subject of returns to scale: if each of the inputs in a productive process (a production function) is increased by the same percentage, does the output increase by exactly that percentage, or by a bit more (increasing returns to scale, commonly shortened to "scale economies") or a bit less (decreasing returns to scale)? Most heuristic treatments of the economics of production deal with the case of constant returns to scale because of its greater simplicity than either type of nonconstant returns. However, although it was implicit in that discussion, we did not emphasize the fact that constant returns to scale in production eliminate virtually all rationale for large-scale production when they are combined with the circumstance of relatively fine divisibility of inputs. When constant returns to scale prevail, we would expect to find production pretty well dispersed over a landscape, as is the case with much agriculture. Consequently, the sheer existence of cities, with their clustering of production and consumption, is basis for suspicion that there would be something empirically wrong about

assuming constant returns to scale, or perfect divisibility of inputs, or both, when we think about cities.

The scale economies associated with production in cities have been found to fall primarily into two principal types of “agglomeration economies,” called localization economies and urbanization economies. They are different mechanisms, as we’ll show. Think of the production function of some firm i in industry j , $q_{ij} = f_{ij}(n_{ij}, k_{ji})$, where $Q_j = \sum_i q_{ij}$ is the total output of the industry of which this firm is a member, and $N_j = \sum_i n_{ij}$ is the total employment in that industry. As we’ve written the production function, it could be of constant returns to scale, and in fact, let’s assume that at the level of the firm, it is. When production in an industry is subject to localization economies, the production of any one firm depends not only on the levels of its own inputs, but on the size of the industry clustered in the same small area. We can represent this dependency formulaically as $q_{ij} = g(Q_j) \cdot f_{ij}(n_{ij}, k_{ji})$, or $q_{ij} = g(N_j) \cdot f_{ij}(n_{ij}, k_{ji})$, where $\Delta g / \Delta Q_j$, $\Delta g / \Delta N_j > 0$, and the subscript j indicates that we are referring specifically to industry j . The output or employment of industry k , different from industry j , has no such effect on the output of industry j in this case although there is no reason to exclude such possibilities, as we note shortly below. This particular agglomeration economy can operate through several nonmutually exclusive mechanisms. This type of economy is external to the firm but internal to the industry; it depends on the scale of activity in the industry, not in an individual firm.

The first mechanism underlying localization economies involves the benefits of labor pooling. Skilled labor has access to a variety of employment opportunities, and firms with demands for skilled labor have comparable access to a pool of such skilled workers. Second, the firms may find scale economies in the provision of intermediate inputs that are common to them all, such as port and transportation facilities, specialized warehousing, energy sources, and so forth. Third, there may be benefits to easy communications among firms, involving information on evolving technology or assessments of demand. Fourth,

there may be opportunities for specialization in the activities of both workers and firms. We can give an idea of the magnitude of this type of scale economy around the contemporary world. Henderson has found that these economies begin to kick in at relatively low levels of industry employment, at least by contemporary standards: 500 employees in the same industry in Brazil, 2000 in the United States; it would not be surprising that such economies would have begun at far lower employment levels in ancient cities. He estimates that a 10% increase in industry employment generates a 1 to 2% increase in output by a firm for the same level of inputs. These scale economies at present tend to be neutral between labor and capital – that is, they raise the marginal products of labor and capital in about the same proportion – although they tend to use high-skilled relative to low-skilled labor. He found these localization economies to be generally unaffected by employment in other industries (Henderson 1988, 224).⁶

Urbanization economies can be represented by $q_{ij} = f_{ij}(P)(n_{ij}, k_{ji})$, where P is the population of the city. The industrial mix represented by the combination of Q_j or N_j does not affect these scale economies. These economies are external to the industry and result from the level of all economic activity in the city. Mechanisms creating urbanization economies include access to a large market, which reduces the need to transport products considerable distances to sell them. The second mechanism is ready access to a wide variety of specialized services which find demands across a variety of industries. Third is the possibility for spillovers of technology and other types of knowledge across industries.

A third form of agglomeration economy is called the interindustry linkage, which is really a crossbred variety of the localization economy. Firm-level productivities in some industries may be enhanced by the proximate location of firms in related but different industries. The principal mechanism for this type of economy is the elimination of or drastic reduction in transportation costs of intermediate inputs. Various kinds of face-to-face interactions – consultations

regarding tailoring of intermediate inputs to another firm's specifications – can also contribute to agglomeration economies.

If no positive production externalities can be derived from locating two different industries in the same city, the city that houses said industry is likely to specialize, within the limits of what “specialization” means in the context of the urban economy, in that industry. Separation of industries into different cities, in such a case, allows each industry to achieve a greater degree of localization scale economies for a given level of commuting costs of workers, which we discuss in the following subsection. Thus, a high degree of specialization within some type of industry in a city is evidence of the weakness of urbanization economies and interindustry linkages relative to localization economies – testable hypotheses.

12.2.2 Externalities

We've already noted the ability to exchange views and ideas with people of different origins as one type of externality that cities confer. Clearly the agglomeration economies of the previous subsection are other types of positive externalities. Cities also generate diseconomies, or negative externalities, as they increase in size: congestion, pollution, crime, and various forms of social conflict as people of different ways of thinking are thrown into closer proximity. The threat of fire and disease also accompany the increasing density of structures and population associated with larger city size.

To the extent that workers live and work at separate locations, a larger city population generally will impose higher commuting costs on at least some workers as their home and work places tend to get separated more. Some additional population to a city will tend to impose longer average commuting trips on previous residents, not necessarily immediately of course, but after both firms and residences have had time to adjust to the price changes that accompany greater population density and higher demands for desirable locations within a city. Since streets are typically common property

resources, they will be subject to congestion, particularly at peak commuting times of the day and week.

At the margin of optimal city size, the additional benefits of agglomeration should just offset the marginal costs of the negative externalities of additional size. There are several decentralized mechanisms by which this marginal balancing occurs. “Decentralized” is important because it means that there doesn't have to be some central planner who recognizes, measures, and regulates these externalities. The mechanisms operate at the level of the individual firm and city resident in the forms of land rents and nominal wages (as contrasted to real wages, which should be equated between city and countryside and among cities).

12.2.3 Types of production

An important distinction between types of goods produced in cities is that between goods that can be traded between cities readily, albeit at some bearable transportation cost, and goods that, for all practical purposes, can't be traded, but rather have to be consumed “in place.” Housing is the prime, urban nontradable good: once built, it is prohibitively expensive to move a house or some other building. Buildings account for the majority of occupied space in the urban landscape, and the industries that construct, maintain, and repair buildings represent an important share of activity in most economies, whether those activities are contained in distinguishable industries or simply represent how individuals allocate their time across their activities. Other nontradables are services: religious services, personal services such as barbering and routine medical treatment, gardening, laundering, repairs of various sorts, retailing, innkeeping, warehousing, and entertainment. Empirically, at least 50 to 60% of a city's labor force works in nontradables. However not all services are nontraded, for example, tourism and more specialized medical treatments.

There are several important characteristics of nontradables. Some of them tend to use relatively high proportions of land – relative

to tradables, although that is not always true. For most nontradables, the consumer has to be at the site of the production: either you go to the barber or the barber goes to you, but when the haircut is produced, the barber and the customer must be at the same location. The same goes for most other services. In the contemporary economy, technology has allowed some services to be provided to spatially separated clients, but this is a recent development that need not obscure this important economic characteristic of most services.

When we speak of a city specializing in the production of some traded good, we are abstracting from the large share of the work force occupied in nontradables. Consequently, accounting for 20 or 30% of a city's work force in a single industry would represent a striking degree of specialization – specialization within the export sector. Again, speaking empirically, small and medium-size cities tend to either specialize in one type of manufactured good or produce services for a local rural hinterland. Larger cities have more diversified production and employment.

The tradable goods sector of a city has been called its “economic base” in a model of a city's economy called economic base theory. Economic base theory developed out of some of the concepts of Keynesian macroeconomics in its use of multipliers to represent the impact of a change in one component of spending on other parts of the economy. The nontradables sector is called the “nonbasic sector” in economic base theory, which as you can see gives a sort of primacy to one type of activity over another which is warranted more by one's preference of subject matter than by priority of either type of production. The nonbasic sector “serves” the basic sector (the nonbasic sector is to a great extent, but not exclusively, the service sector), and as the demand for the basic sector's product or products grows, employment in the nonbasic sector grows in some proportion to basic employment growth – roughly in proportion to the number of basic-sector employees a “typical” nonbasic sector employee can serve. The model emphasizes the demand side of the city's

employment at the expense of forces that tend to boost supply. We'll use a more sophisticated model of the external linkages of cities in section 12.5, one that offers a more balanced treatment of the relationships between tradables (“export” or “basic”) and nontradables (“nonbasic” or “services”) production.

Contemporary economics has not escaped the taxonomic temptation, but classifications of cities tend to be avowedly partial, and alternative classification schemes need not classify cities identically. One recent, popular and influential classification of contemporary American cities relies on characterizations of the economic activities in different types of cities. The functional typology of Noyelle and Stanback (1983) divides U.S. cities into five primary categories: diversified service centers, specialized service centers, consumer-oriented centers, and production centers. Before you say, “Aha! There are still ‘consumer cities’ and ‘producer cities’ in contemporary thought!” let me say a few words about these last two types of cities. The Noyelle–Stanback taxonomy includes two types of consumer-oriented city, residential and resort-retirement. The former is essentially a large suburb, and the latter has its consumption activities financed to a large degree by savings of one sort or another. The production centers in their taxonomy are divided into production, industrial-military, and mining-industrial, while many cities traditionally famous for their manufactures, such as Detroit, Pittsburgh, and Toledo, fall into the “functional-nodal” subtaxa of specialized service centers. The other subspecies of the specialized service center are government-and-education and education-and-manufacturing. This is a statistically derived grouping based on the mix of economic activity as measured by employment shares.⁷ We need not think seriously of its application directly to cities of the ancient Mediterranean region, but the concept of a multivariate classification scheme based on the basic / nonbasic or tradable/nontradable distinction is worth retaining. Comparable information on the ancient cities may not be obtainable, but the embodied idea that there are several functional types of city at a given time is superior to the blanket classification of the

city sui generis of a period as being of one type or another.

12.3 Housing

Just as land is a sufficiently distinctive type of capital to warrant separate consideration as a factor of production, housing is a peculiar enough (durable) consumption good to justify its own discussion. First we describe the characteristics that make housing such a special consumption good, then turn to separate, brief discussions of housing supply and demand.

12.3.1 *The Special Characteristics of Housing*

First, the importance of housing in a consumer's array of consumption goods is striking. As shelter, it is a necessity, and nearly all people consume it. It is also expensive, and for most households it is the single largest item of budgeted consumption. While contemporary figures from the United States cannot be transferred unhesitatingly to the societies of the ancient Mediterranean, 36% of all households in central cities spent more than 36% of their household income on housing in 1991; among renters, that figure was 48%. The median ratio of housing cost to income in central cities was 24% (McDonald 1997, 205). Typically the value of a residence is equivalent to several years' income of the resident household. For most owning households, the dwelling unit is the single largest asset in their portfolio. Residential structures accounted for 40% of net national wealth of the United States in 1958 (Goldsmith 1962, 177, Table A-35). As a consequence of the costliness of housing, many consumers rent their housing services, and in contemporary, relatively well developed capital markets, most owners pay for their units over lengthy time periods. In developing countries where building codes are not as strict (or as strictly enforced), houses may be built gradually over time as households acquire construction materials. As another consequence of housing's expense, most households can afford to live in

only one place at a time. Housing surely was a correspondingly important item in the ancient urban resident's budget.

The second outstanding characteristic of housing as a consumption good is its durability. In the contemporary United States, housing is relatively short-lived at around 70 years. European housing frequently has lasted several hundred years. The apparent lengths of some pottery periods in the ancient Aegean seem to imply places and times of considerable durability of residential structures although the fires in Rome were consistent with shorter expected half-lives. Structures of different ages can offer competitive housing services.

Third among housing's prominent characteristics is its multidimensionality and heterogeneity. Housing has nearly infinite variety when we take into consideration all the dimensions on which units can vary: number of rooms, floor space, general layout, construction material, architectural characteristics, number and placement of windows, ventilation and heating systems, plumbing characteristics, number of stories, the condition of the structure at a given time, its location relative to various sites to which access is valued, its age, and even its neighborhood amenities. Housing's durability contributes to the heterogeneity of the housing stock of a city at any given time. There is corresponding heterogeneity in taste for housing, and there is substantial substitutability among structural characteristics. This high dimensionality of housing contributes to the difficulty of constructing compact measures of it. Blanton (1994) offers a contemporary, international perspective on the variability of physical housing characteristics as well as their relations to their societies' practices.

Fourth among housing's special characteristics is its spatial fixity. Housing must be consumed where it is built because moving it is prohibitively expensive. Related to its spatial fixity is the fact that the consumer of a house also consumes the amenities of the neighborhood in which the unit is located, whether they are considered positive or negative.

As a consequence of the third and fourth characteristics of housing, the market for housing

may be “thin.” That is, there may be few units that exactly match the demand of a household and correspondingly few demanders an existing unit exactly suits. A unit may be the right size, have the right configuration of rooms, and an agreeable layout but be located farther from some point to which the demander would like good access or be in a neighborhood with characteristics the buyer does not like (for example, too many Hittites, a pigsty or bronze smelting operation next door, and so forth). These consequences of multidimensionality and spatial fixity of housing give both the seller and the potential buyer of a unit some bargaining power in determining either a sale price or a rental.

12.3.2 *Housing supply*

Four production processes contribute to housing supply: construction, maintenance, rehabilitation, and conversion. The archaeological record contains evidence of all four of these processes. At least in contemporary Western cities, new construction can add from 2 to 5% to the existing housing stock in any given year. This construction rate does not imply that the housing stock grows by that rate annually, because some housing is taken out of the market at the same time, through abandonment, demolition, and accidental destruction (generally fire).

Of course, rebuilding after the destructions of World War II, as well as reconstructions of cities after fires and earthquakes in recent times – for example, Chicago in 1871 and San Francisco in 1906 – demonstrates the surge capacity of economies for emergency reconstruction of urban infrastructure, at much faster annual rates. This observation bears consideration when contemplating the repeated destructions of cities in the eastern end of the Mediterranean and the Aegean in antiquity, by both earthquakes and human agency, either accidental fires or warfare. However, in these modern reconstruction efforts, assistance has come from outside the destroyed areas. The evidence from antiquity of alternations between postdestruction resiliency and stagnation or abandonment may be indirect evidence of the ability of the affected cities to obtain outside assistance,

either from the city's own natural hinterland of countryside and smaller towns or from neighboring political units. Alternatively, their locational advantages also may have been eroded or destroyed.

Maintenance is the gradual application of capital and labor to an existing structure with fixed floor space and configuration, with the goal of maintaining the quality or quantity of service flows at a given level, or alternatively of slowing the rate of deterioration in those flows. Rehabilitation is similar to maintenance but involves the sudden application of capital to produce a discontinuous change in service quality. Maintenance of housing, as with other capital equipment, is a means of adjusting an existing capital stock to an alternative desired level. If the service flows from a structure at a particular time and location in a city are higher than the flows that are demanded, reducing the flow of maintenance into the structure will reduce the flows. If at some later time, a higher rate of service flows is demanded from the same structure, a sharply increased level of maintenance, or rehabilitation (which is really a variant of maintenance), can increase the capital stock in the unit and raise the service flows. Of course, the relevant question is whether the condition of the housing unit at the time more housing service flows are demanded from that location will permit the rehabilitation required to restore the flows to the desired level at a cost equal to or less than the valuation of the higher level of services (that is, can the rehabilitation be profitable, either to separate contractors and residents or to a contractor and resident wrapped up in the same person?). If the condition of the structure is too run down for rehabilitation to restore its service flows to the desired level, the building very well may not be demolished to permit construction of a new structure that will supply the demanded level of housing flows simply because demolition is expensive. If the demand for a structure in its current condition (that is, for a level of housing flows that it can provide) is less than the cost of maintenance to keep it at that condition, the unit is likely to be abandoned by the owner. Whether it is reoccupied by squatters depends on income levels and the legal status of squatting. There

certainly seems to be evidence of such occupation in and around some of the former late palace at Knossos on Crete.

Conversion involves changing the size or configuration of a building at a fixed location and structural density in the neighborhood. As with rehabilitation, it generally occurs discontinuously. The thinness of the market for housing in a city, noted above, suggests that a good deal of conversion activity will take place when a unit changes owners (assuming that renters are restricted from making such alterations).

A concomitant of durability and costliness of demolition is the high degree of irreversibility of an investment in housing. The scrap value of components typically is low because many cannot be removed without damaging them, and by the time many components are installed, they have been sawed or hammered into quite specialized (that is, nonstandardized) shapes and sizes. However, in an era that was short on standardization, the retrievability of odd building parts might have been quite valuable, as evidenced by the widespread reuse of stone from buildings over periods of centuries. But in general, the elements of irreversibility that do exist will lead suppliers of housing to be cautious in their responses to what might prove to be short-term fluctuations in demand. That said, evidence exists of widespread remodeling of ancient urban dwelling structures, as reported by Stöger (2011, Chapter 5) for *Insula IV* at Ostia.

Altogether, in normal economic times, the supply of housing responds sluggishly to changes in demand. If demand falls, suppliers slack off on maintenance and the quantity of effective housing at certain locations declines over the next few periods. If demand rises sharply, the price of existing housing is likely to rise before sufficient new units can be constructed to meet the increased demand. Although the long-run supply elasticity of housing is quite high (it may approach perfect elasticity, or positive infinity), the short-run supply elasticity is quite low.

12.3.3 *Housing demand*

The current best estimates of the demand for housing in the contemporary United States give

an income elasticity of demand in the range of 0.8 to 1.0 and a price elasticity in the range of -0.5 to -0.8 (0.5 and -0.4 in a recent estimate among British owners) (Ermisch *et al.* 1996, 66–67, 82). Once again, we cannot transfer these magnitudes directly to the ancient world, but neither should students of ancient housing ignore these general ranges. The demand for housing is likely to be both price- and income-inelastic, so a doubling of prosperity should not lead to a doubling of housing sizes, with housing prices constant. As we will see in section 12.4, the size of a city will influence the price of a house of a given size, holding constant the real income of the populations of the two cities. The mechanism is the increase in population density and the corresponding bidding up of land prices, which form part of housing prices. So in bigger cities, we would expect to see a higher average house price than in a smaller city at corresponding locations within the two cities. If the housing price in the larger city was double that in the smaller one (a fact that probably will remain unobservable to us but would have affected behavior nonetheless), the average size of housing unit in the larger city, holding household real incomes constant would have been in the neighborhood of 60 to 80% of that of the typical housing unit in the smaller city. At least, this is a reasonable range to start testing one's guessing with remains that *are* observable.

In contemporary cities, location has a strong influence on a household's housing demand because of both employment location and neighborhood amenities. If most people worked at their residence, the neighborhood amenities would still be influential, as well as the price effects on locations contributed by households who did have a strong demand for access to some fixed locations in the city.

Household composition has a strong effect on housing demand, which implies that one's stage in life cycle affects demand. Nonetheless, just as housing supply adjusts slowly to changes in demand, households don't adjust their choice of housing unit rapidly as their own circumstances change (although conversions may be performed over the life cycle). People expect to remain in their houses for considerable lengths of time (whether for several years or for several decades),

so one's forecast of the future is an important consideration in housing choice. This in turn implies that decisions about housing are both consumption and asset portfolio decisions.

12.4 Urban Spatial Structure

The density of population and structures inside cities typically is not evenly distributed but has a regular falling off away from some central point of attraction. In recent and many contemporary cities, the geographical center of the city, generally called the central business district (CBD), has been the principal such focal point. Typically these focal points within cities have started as dock or wharf areas or junctions of major communication systems. Many agents have found it valuable to be located near these sites and they accordingly have bid up land rents in the vicinity along the lines the Thünen model predicts for such a fixed point of attraction. As economic activity continued to develop in these areas they became important areas of employment, residential proximity to which was valuable for employees who had to travel from home to work each day. Central city land prices continued to be bid upward. As construction technology permitted taller structures and urban economies attracted work suitable to such tall buildings, a single unit of land could yield yet higher rent. As residences began to surround these focal areas relatively symmetrically, these central regions found their desirability redoubled as sites most accessible on average to work forces living in the city.

Not all types of production have found the high capital-land and labor-land ratios suitable to high-rent, central-city sites suitable to their technical operations. Moving to, or locating originally at some distance from the CBD, such activities have initiated the development of secondary centers of attraction for residential and other business purposes, yielding corresponding local peaks in land values outside the CBD. Yet subsequent development of shopping districts closer to residential areas distant from the CBD has reinforced this trend to development of secondary

centers, sometimes developing independently and rivaling the original CBD. Wiegand (1987) models the development of such a secondary center in a monocentric model.

This thumbnail sketch of the development of the current spatial structure of the typical city in the contemporary world relies on many technologies that are relatively new. Mass transportation, even horse drawn, encouraged the spatial separation of home and workplace to an extent not possible in purely pedestrian cities. Steam- and fossil-powered transportation systems reinforced this trend, as did steel-frame building construction technology. In the absence of these essential traits of modernity, can this model of the internal spatial structure of cities help inform us about ancient cities in the Mediterranean basin? In the first place, there are really no alternatives to guide us, so in the best tradition of making do with what's available, we shall trot it out and do our best with it. Second, the situation may not be as bad as all that, because many of the forces operating at the level of individual consumers and producers to generate the aggregate spatial density patterns of contemporary cities would have operated in antiquity, although with less capacity to exaggerate price differentials between locations. Location near some points within ancient cities, whether for residential, business, administrative, or religious purposes, would have been particularly desirable; in fact, activities across these categories may have competed with one another for sites in the same vicinity.⁸ We may have little information on what activities found proximate location to one another valuable, even if we are confident that the majority of work occurred close to, or at, home, depressing the effectiveness of the home-workplace attraction to create land rent differentials.⁹

With this apology we'll turn to the monocentric city model. We follow that exposition with brief modifications to account for locations of people with different economic characteristics; for some, but not all, people working at home; and for the endogenous formation of central land rent peaks in contrast with their exogenous specification in the simplest versions of the model.

12.4.1 *The monocentric city model*

This model posits a single, central place of employment of all residents in the city and uses this spatial separation to generate a land rent gradient and a residential population density gradient and predictions about city size (radius) and total population. Starting to work with a multi-nucleated city simply clutters up the analysis without adding any additional insights worth the trouble. We'll develop the model in two distinct components: the consumer's demand for housing and the production decisions of the suppliers of housing. From the consumer's side of the problem we get the spatial gradients of housing prices and housing consumption per person (defined as floor space per person; the inverse of this, of course, gives us population density) from the center of the city to the edge. From the producer's side of the problem we get the spatial gradients of building density and land rent. Then, depending on whether the city is "open"¹⁰ or "closed,"¹¹ by various modeling devices we can either add population until we hit the exogenous utility level and discover the corresponding city radius or fill the city with the exogenous population and discover the corresponding utility level and city radius.

Thus the model determines the spatial gradients of land rent, housing price, housing consumption (the inverse of which is residential population density), structural density; the values of those variables at the city center; and the radius of the city. The open version of the model also determines city population, while the closed version determines the uniform utility of the city residents. The exogenous variables used to determine the values of the endogenous variables are agricultural land rent, per capita or household income, and the transportation cost of getting back and forth to work; utility is exogenous in the open model, and city population in the closed model. The interpretation of income in the open model needs care, since utility is fixed, regardless what we say happens to income; in this case income is a nominal wage rather than what we'd think of as "real income."

Let's start with the consumer's problem of choosing a location in the city (k , defined as the distance along any ray from the CBD) and a quantity of housing services at that location, $q(k)$, measured as square feet of floor space. The following exposition of the monocentric city model is adapted from Henderson (1985, Chapter 1), Straszheim (1987), Brueckner (1987), Turnbull (1995, Chapter 2), and Wheaton (1974). The only expenditure on transportation to and from work at the center of the city is the time spent traveling – effectively walking. Assuming that everybody works the same number of hours, T , this time comes out of leisure, $e(k)$, the notation of which indicates that the amount of leisure one enjoys depends on one's location relative to work, k . Then leisure at any location is $e(k) = T - tk$, where t is the time spent walking one (round-trip) unit of distance. Naturally, we begin with the further assumption that all individuals have the same taste and real income. The prototypical individual's utility function is $u(k) = u[z(k), q(k), e(k)]$, where z is a composite commodity representing all other consumption besides housing; making z -consumption a function of k is our first hint that there will be a tradeoff between housing consumption and the consumption of all other goods; the fact that both housing and z -consumption are functions of location indicates that the tradeoff will differ across locations. The consumer's budget constraint is $y = p_z z + p_h(k)q(k)$, which simply says that she allocates her entire income, y , between consumption of the composite commodity and housing (p_h is the price of housing per unit of floor space). Her time constraint is simply a rearrangement of the definition of leisure: $T = e(k) + tk$.

The consumer's maximization problem is $\text{Max}_{\{z, q, e, k\}} \mathcal{L} = u[z(k), q(k), e(k)] + \lambda[y - p_z z(k) - p_h(k)q(k)] + \lambda[T - e(k) - tk]$, which she accomplishes by jointly choosing the quantities of the composite good, housing, leisure and a residential location. As you no doubt recall by now, we as analysts reproduce this process of searching for the optimal values of these consumption flows and location by "taking the first-order conditions." One of the outcomes of finding the first-order conditions is a demand

function for each item of consumption. For housing, the demand function can be expressed as $q(k) = q[y, p_h(k), p_z, e(k)]$, which simply says that the demand for a flow of housing services at any location k is a function of the consumer's income, the price of a unit of housing at location k , the spatially invariant price of the composite commodity, and the quantity of leisure consumed at location k . This is just like the demand functions we derived in Chapter 10, where we developed the theory of the joint determination of commodity consumption and leisure consumption, or equivalently, labor supply. In this case, while we've held constant the part of labor supply actually spent at work, we've allowed the time spent getting to work to be part of the choice of the full amount of time devoted directly or indirectly to working.

From the first-order conditions for leisure and location, after making a few manipulations, we obtain an important spatial equilibrium condition: $q(k) \cdot [\Delta p_h(k)/\Delta k] = -p_e(k)t$, where p_e is the "price" of leisure expressed in terms comparable to income ($p_e(k) = -[\Delta u/\Delta e(k)]/\lambda$, where λ is the Lagrange multiplier from the budget constraint in the constrained maximization problem, which has the interpretation of the marginal valuation of an additional unit of income). The price of leisure is the marginal cost of travel time. This condition says that at her optimal location, the consumer can't improve her welfare by moving: if she moves a short distance k toward the work site to save on travel costs, she will have to spend exactly the value of what she saved on transportation on increased housing costs. We know that the marginal utility of leisure is positive $[\Delta u/\Delta e(k) > 0]$, so p_e being positive implies that the spatial gradient of the housing price is negative $[\Delta p_h(k)/\Delta k < 0]$. Conversely, if our consumer were to move farther from work to save on housing expenses, she'd find that her valuation of her foregone leisure spent in traveling the extra distance would exactly offset the gain in utility she experienced from the additional (and cheaper per unit) housing consumption. Rather than representing a bizarre form of frustration devised by economic theorists, this condition simply says that in an equilibrium, the consumer chooses a combination of consumption flows that she can't improve upon, given prevailing prices and her income, which is the standard equilibrium condition for consumption.

We show these choices diagrammatically in three steps in Figure 12.1. First, in Figure 12.1(a), we lay out the indifference curve from the individual's indirect utility function. Remember that we're working with the open city model in which utility is exogenous. This is in contrast to the way we usually work with indifference curves to analyze other consumption behavior: we give the consumer a fixed budget constraint and let her reach the highest level of utility she can with

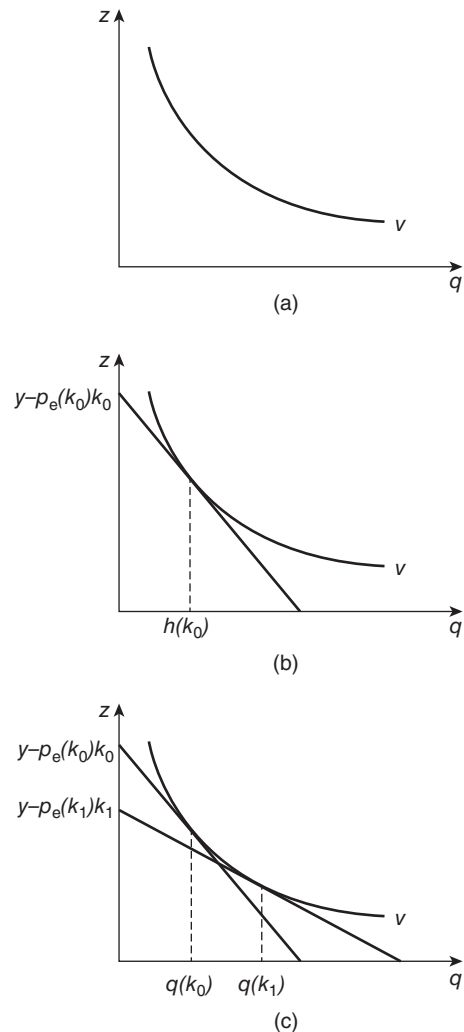


Figure 12.1 (a) An individual's indifference curve for location versus housing space. (b) This individual's consumption of a composite commodity if residing at a given distance from the city center. (c) Equilibrium choice of housing size and location.

that income. In Figure 12.1(b), we pick a distance from the center of the city, k_0 , and, using the full-income concept, find the maximum amount of the composite commodity the consumer would be willing to consume after paying transportation costs from k_0 to the work site at the city center, $y - p_e(k_0)k_0$. This point nails down the vertical intercept of the budget line. Next we rotate the budget line around its vertical intercept until we find its tangency with the fixed indifference curve. The negative of the slope of the budget line when this tangency is found is the relative price of housing at location k_0 , at which price our consumer wishes to consume $q(k_0)$ floor space of housing. Now, let's see what happens when we take this consumer to a more distant location, k_1 . In Figure 12.1(c), we reproduce the original choice of $[z(k_0), q(k_0)]$ at location k_0 , but we've added the constraints facing consumption at location k_1 . With higher transportation costs at k_1 , the maximum consumption of z is lower than at k_0 : at $y - p_e(k_1)k_1$. Anchoring the vertical intercept of the net-of-transport-cost budget line at that point, we swivel the budget line till we reach the tangency with the same indifference curve the consumer faced at location k_0 , which occurs at the level of housing consumption denoted $q(k_1)$, which is greater than $q(k_0)$. In moving the consumer farther from the city center, from a residential location on more expensive land to one on cheaper land, she chooses a larger quantity of floor space and less of the composite commodity, but she keeps the same level of real income identified by the constant indifference curve. The flatter slope of the budget line coming out of vertical intercept $y - p_e(k_1)k_1$ represents a lower price of land, $p_h(k_1) < p_h(k_0)$.

Now let's turn to defining further the spatial gradient of housing prices. Remember that our city is placed in what we could consider an agricultural plain: the city has to out-compete agriculture for land on which to build urban housing. This is equivalent to the condition that at the edge of the city, the rent that urban uses of land must generate is equal to the alternative rent the same unit of land could produce in agriculture. Figure 12.2 shows such a gradient from the CBD to the point at which the price of housing on urban land equals the price a unit of housing would cost if it were built on agricultural land (k^*). Now, let's turn to the indirect version of the

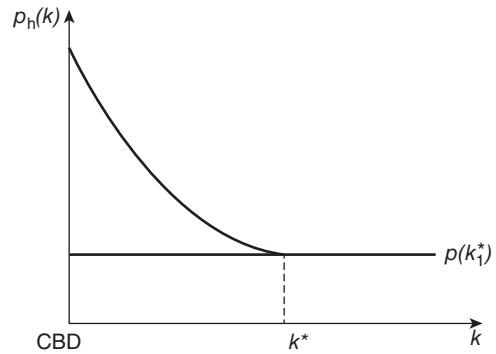


Figure 12.2 Spatial gradient of housing prices moving away from a city center.

utility function with which we began. Remember that the indirect utility function expresses the same information that the direct utility function does, but does so in terms of income, prices, and leisure: $v(k) = v[y, p_z, p_h(k), e(k)]$, in which $\Delta v / \Delta y > 0$ (more income gives more utility, given constant prices), $\Delta v / \Delta p_i < 0$ (higher prices of any consumption good reduces utility, given a constant income), and $\Delta v / \Delta e > 0$ (more leisure yields higher utility, with prices and income held constant). We know p_z (it's exogenous to the model). From the supply side of the monocentric city model, which we'll show directly below, we can find the price of housing at k^* from the production function for housing and its dual form, the housing unit cost function, in terms of the price of capital and the rent of agricultural land, both of which we take as given. Then indirect utility at k^* is defined by leisure at k^* , given the commuting time at k^* , tk^* , which in turn defines overall utility as $v(k^*)$ throughout the city. Then, at any location in the city, from the known values of $v(k^*)$ and $e(k^*)$, we can solve for the only remaining unknown in the indirect utility function, $p_h(k)$, which is the height of the housing price gradient at location k . Consequently, the price of housing at any location is a function of household income, commodity prices, the known housing price at the edge of the city, the city radius, transportation technology, and the amount of time available to be divided between leisure and travel: $p_h(k) = p_h(y, p_z, p_h(k^*), k^*, k; t, T)$, in which the presence of both k^* and k as arguments in the housing price function indicates that the

housing price varies according to location k within the city but at any given location k , there is an independent influence of the radius of the city. These two factors have opposite influences: $\Delta p_h(k)/\Delta k^* > 0$, indicating that, other things being the same, a larger city will have higher housing prices at any given distance from the city center; and $\Delta p_h(k)/\Delta k$, which indicates that regardless of city size, housing located farther from the city center will be cheaper than housing closer to the center. We still need to determine the value of k^* , which is accomplished by bringing total city population into the model. We'll do that after introducing the supply side of urban housing.

The suppliers of housing produce housing with capital and land according to some production function $h(k) = h[K(k), L(k)]$, where $K(k)$ is spatially mobile capital (bricks, stone, metal, and boards) used at location k and $L(k)$ is land at location k . $\Delta h/\Delta K > 0$ (or in alternative notation, $h_K > 0$) as is usual for production functions (more capital, more house); and $h_{KK} < 0$ because as more capital gets used in a single building on a constant unit of land, we must be talking about a building that gets taller, and the capital is increasingly used to strengthen foundations and lower walls and to build stairways, both of which reduce the area of floor space in which we're measuring housing. The revenue from housing is $p_h(k)h(k)$, and production costs are the capital rental on the capital used, $iK(k)$, and the land rental on land, $r(k)L$; profit then is $\pi(k) = p_h(k)h(k) - iK(k) - r(k)L$. With competitive supply of housing, the retail price of housing will equal its unit cost, and the cost function of housing can be expressed, in parallel fashion to the indirect utility function on the consumer's side of the problem, as $p_h(k) = p[i, r(k)]$.

The spatial variation in housing prices causes spatial variation in the gross revenue from a unit of housing, leading housing suppliers to want to choose profit-maximizing input combinations as well as profit-maximizing locations. When a producer has found a profit-maximizing location, $\Delta\pi/\Delta k = 0$ and $h(k)[\Delta p_h(k)/\Delta k] = L(k)[\Delta r(k)/\Delta k]$, which says that when a housing producer moves his construction site a bit, his land costs exactly offset the gain he can make in gross revenue from the housing. We can rearrange this equilibrium condition just a bit to see

the relationship between the housing price gradient and the land rental gradient: $\Delta r(k)/\Delta k = [h(k)/L(k)][\Delta p_h(k)/\Delta k]$. We already know that $\Delta p_h(k)/\Delta k < 0$, so the land rent gradient is negative as well. Changes in land rents across small changes in location exactly equal changes in housing costs, and we already know that the latter also equals the value of the loss in consumers' leisure of the corresponding locational changes. Thus, $L(k)[\Delta r(k)/\Delta k] = h(k)[\Delta p_h(k)/\Delta k]$, and $q(k)[\Delta p_h(k)/\Delta k] = -p_e(k)t$. Remember that $h(k)$ is built floor space at location k while $q(k)$ is floor space per dwelling unit at location k ; floor space divided by floor space per dwelling, assuming one person per household, gives population density at k : $D(k) = h(k)/q(k)$.¹² We know that $\Delta q(k)/\Delta k > 0$, and we'll see below that $\Delta h(k)/\Delta k < 0$, so $\Delta D(k)/\Delta k$, the urban population density gradient, is negative.

Just as we used the indirect utility function to solve for the height of the housing price gradient at any location in the city, we can use the unit housing cost function to solve for the height of the land rent gradient at any location. The rent on a unit of land at location k will be a function of household income, commodity and factor prices, transportation technology, and the total amount of non-work time available to use as leisure: $r(k) = r(y, p_z, p_h(k^*), i, k^*, k; t, T)$. Land rent is clearly a derived demand.

We can rewrite the housing producer's spatial equilibrium condition to reveal more generally the relationship between the housing price gradient and the rent gradient: $\% \Delta r(k) / \% \Delta k = [\% \Delta p_h(k) / \% \Delta k] / s_L$, where s_L is the share of land in housing costs, $r(k)L(k)/p_h(k)h(k)$. Since $s_L < 1$ (and probably around 0.1 to 0.2 regardless of the time or place), housing rents have a magnified effect on land rents. Thus, if we travel some distance across a city, say moving toward the center, and find a 10% rise in housing prices, we could expect a 50% to 100% rise in land rents over the same distance. Another rewriting of the spatial equilibrium condition gives the corresponding relationships between structural density, $K(k)/L(k)$, which we'll define as $S(k)$, and the housing price and land rent gradients: $\% \Delta S(k) / \Delta k = \sigma \% \Delta r(k) / \Delta k = (\sigma / s_L) \% \Delta p_h(k) / \Delta k$, where σ is the elasticity of substitution between land and capital in the production of housing. This set of

relationships says that a 1% increase in housing prices as we approach the center of the city produces a $\sigma/s_L\%$ change in structural density. Thus, if the elasticity of substitution is, say, 1, a 10% increase in housing prices would cause a 50% to 100% increase in structural density over the same distance. Another measure of the intensity of land use is the value of structures per unit of land, $p_h(k)h(k)/L(k)$, which we'll call $V_h(k)$. This measure is related to housing prices as $\% \Delta V(k)/\Delta k = [1 + (\sigma s_K/s_L)]\% \Delta p_h(k)/\Delta k$, where $s_K = 1 - s_L$ is the share of capital in housing production costs. With a 10% land share in housing and a unitary elasticity of substitution, a 10% increase in housing prices would be associated with a doubling in the value of structures per unit of land.

The next task is to find the total population and area of the city. We have a measure of population density at any location, $D(k)$. We have three relationships that will help us solve for these variables. First, the total population at distance k from the center of the city is $2\pi D(k)$, and the aggregate population of the entire city, N , is the sum of these concentric populations from the center of the city to its edge at k^* . So $N = 2\pi \sum_{k=CBD}^{k^*} D(k)$. Second, we know that land rent at the edge of the city is equal to land rent in agriculture: $r(k^*) = r_{agr}$. Third, in the open city model we know that indirect utility at the edge of the city, which we can characterize as a function of income net of transportation costs and land rent (or housing prices) at the city's edge, must remain equal to the exogenous level of utility throughout the city: $v[y - p_e(k^*)k^*, r(k^*)] = \bar{v}$. Simultaneous solution of these three relationships gives us land rent at the center of the city, $r(CBD)$, the city radius, k^* , and the city population, N . This solution

set fills the city with just the population that makes the most distant residents – those located at k^* – travel the distance that will keep their utility at $\bar{v}$; by virtue of the spatial equilibrium condition for housing consumption, all residents will experience just the combination of travel time and housing consumption that keeps them at the same utility level.

Tables 12.1 and 12.2 report the responses of the city's principal structural features under the open and closed assumptions (Wheaton 1979, 111). Comparing the responses of the various gradients to income and transportation costs, you'll see that the closed city has more intricate responses to these parametric changes. This is because the fixed utility level of the open city eliminates the traditional income effect from price changes, leaving only the constant-utility substitution effect and eliminating the possibility of income effects outweighing substitution effects over certain distance zones. Thus, looking at the last two columns of the two tables, which show the responses of structural components to variations in income and transportation costs, in the open city, the distance gradients that comprise city structure move either up or down throughout the city whereas in the closed city, they tend to fall in some parts of the city and rise in others as the gradients rotate rather than simply shift. These tables do not show the influence of capital prices on these gradients because they are impossible to assess unambiguously. A higher capital price will raise the demand for land relative to the demand for capital, and it depresses the demand for housing and consequently the demand for both factors – capital and land. Whether the sum of the influences on the aggregate demand for land is positive or negative – whether k^* expands

Table 12.1 Responses of city structure to changes in parameters, open city.

<i>Effects – changes in:</i>	<i>Causes – increases in:</i>			
	<i>utility</i>	<i>agricultural land rent</i>	<i>income</i>	<i>transportation costs</i>
housing prices	down	invariant	up	down
housing consumption	up	invariant	down	up
land rent	down	invariant	up	down
building density	down	invariant	up	down
population	down	down	up	down
edge of city	down	down	up	down

Table 12.2 Responses of city structure to changes in parameters, closed city.

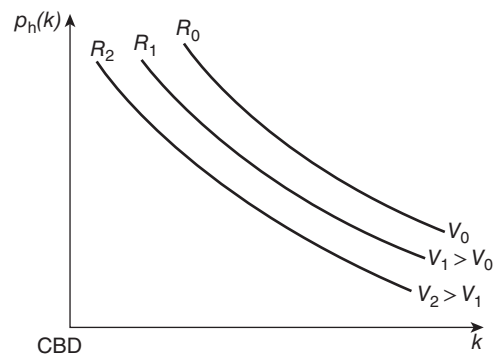
<i>Effects – changes in:</i>	<i>Causes – increases in:</i>			
	<i>population</i>	<i>agricultural land rent</i>	<i>income</i>	<i>transportation costs</i>
housing prices	up	up	down toward center, up toward edge (gradient rotates)	up toward center, down toward edge (gradient rotates)
housing consumption	down	down	rises wherever housing price falls	rises wherever housing price falls
land rent	up	up	down toward center, up toward edge (gradient rotates)	up toward center, down toward edge (gradient rotates)
building density	up	up	down toward center, up toward edge (gradient rotates)	up toward center, down toward edge (gradient rotates)
edge of city	up	down	up	down
utility	down	down	up	down

or contracts – remains an empirical question, or at least one dependent on the relative values of a number of parameters in demand and production functions. There is no general answer to how city size responds to the capital rental rate, only particular cases.

12.4.2 Multiple categories of residents

The exposition of the monocentric city model here has used only one category of consumer. If we wanted to consider two categories of consumer distinguished by, say, income, we would face the problem of assigning the groups to mutually exclusive locations. If the two groups had identical preferences and differed only in income, one group would outbid the other for locations over some zone in the city; their residential locations would not be mixed randomly over space.

We introduced the bid-rent function in Chapter 11 for the Thünen land-use model in its agricultural applications. It is useful to think of the bid-rent function as deriving from the indirect utility function: when all other prices are known at each location throughout the city, the height of the bid-rent curve at any location is the housing price that a consumer would have to pay to keep her utility at the spatially constant level. Figure 12.3 shows a family of bid-rent curves for three different utility levels. The higher utility levels are associated with lower bid-rent curves

**Figure 12.3** Family of bid-rent curves representing different utility levels.

on the reasoning that the less rent you have to offer, the better off you'll be. (Remember that with the indirect utility function, $\Delta V / \Delta p_i < 0$, for any good i .)

The bid-rent apparatus is a useful tool with which to study this type of intracity location problem of “who lives where.” We can get a quick formulation of a bid-rent function by rearranging the budget constraint to solve for rent per unit of land. Returning to our two-group case, we would set the bid rents of the prototypical member of each group equal to each other at the common location k , as Figure 12.4 shows. The group with the steeper bid-rent function at that location will claim the inner zone of the city for its residences

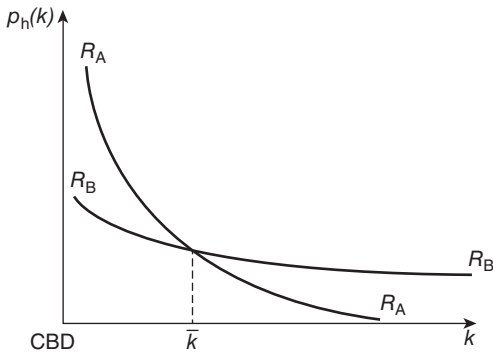


Figure 12.4 Different bid-rent functions of different groups of people.

while the group with the flatter function at $\bar{k}$ will claim the outer zone.

What characteristics can produce a steeper or flatter bid-rent function? Under many conditions, a larger income elasticity of demand for housing will tend to flatten the bid-rent function. Different groups may face different opportunities and costs for transportation – some may ride while others walk, frequently depending on income – and a comparative advantage in traveling can flatten the slope of the bid-rent function. More intricate, multiperson budget constraints open the range of possibilities for influences on bid-rent slopes.

Evidence exists that these forces operated in antiquity. Meiggs (1973, 236) cites Vitruvius' observation that population increases at Rome raised spatial residential densities in the fashion the monocentric city model has produced: "... It is necessary to provide dwellings without number. Therefore... necessity has driven the Romans to have recourse to building high..." (ii.8.17). Augustus imposed a 70-foot building height limit, which Trajan lowered to 60 feet, which still would have allowed four stories comfortably, and possibly a short fifth story. As Meiggs describes the process for Ostia, "When space was scarce and ground rents high the natural solution was for several families to live together in a single building and make the maximum use of the building plot by extending upwards rather than outwards" (Meiggs 1973, 236). The two principal types of urban residential structures in Roman Italy were the *insula*, which was a large, tall block building divided into separate apartments separately let, and the *domus*, a house designed primarily for a

single family (the equivalent of what contemporary censuses describe as "detached single-family dwellings"). The fourth-century C.E. regionary catalogues enumerate 46 000 *insulae* and 1790 *domi* in Rome, evidence of the dense residential occupation of a large city. Cicero, in *Pro Caelio*, 17, noted that a wealthy young friend of his had rented rooms in Rome for 10 000 sesterces (Meiggs 1973, 237–238). Turning to Ostia, Meiggs notes that whereas there was no sign that space was unduly scarce in late Republican or early Principate times (walls from those times not appearing to be intended to carry heavy weights), by the end of the second century C.E., Ostia's architecture had been transformed, and housing conditions had been revolutionized by the replacement of most of the independent atrium houses with tall apartment blocks during Hadrian's reign: even the rich were moving into apartments by the second century C.E. "As population increased [by Hadrian's principate] and the value of land rose the temptation to leave open spaces in the busier parts of the town was easily resisted" (Meiggs 1973, 131–141; quotation from 141).

Nonetheless, it is apparent that the monocentric model sorts different incomes horizontally rather than vertically as was noted for Rome and Pompeii in Chapter 3. The culprit for this prediction is the central workplace assumption, along with the model's lack of distinction of space within dwelling units. Clearly, careful modeling of ancient circumstances could remedy this predictive deficiency. The following subsection takes the first step, modifying the workplace assumption.

12.4.3 Working at home

A central business district where all city residents work is a convenient fiction with which to model the contemporary city, in which so many workers do work away from their residences and such a great concentration of employment raises land values far above their values elsewhere in the city at some location that residents commonly call "downtown." We've seen the spatial equilibrium of the housing market under such conditions: at a resident's optimal residential location, if he moves a bit farther from the city center to enjoy lower land prices, he must pay additional

transportation costs fully equivalent to the saving in housing costs. What kind of mechanism would equilibrate the market for residential space if people didn't have to travel to work?

Imagine a city in which a large percentage of workers work at home yet there remains a focal point in the city toward which land values rise – say a wharf area in a port city, with adjacent warehouses, along with some government offices. Demand for access to the dock and government office area by a number of activities dispersed throughout the city would yield a negatively sloped bid-rent curve throughout the mixed residential/workshop area as we move away from the central area of higher land values. In this setting, the typical consumer's utility function is $u(k) = u[z(k), q(k)]$, which he maximizes subject to the budget constraint $y(k) = p_z z(k) + p_h(k) q(k)$. The following exposition is adapted from Muth (1969, 42–43). Whether we use the open- or closed-city assumption, utility should be spatially invariant across the city; since some people must live on land that has higher value because of its proximity to the port facilities but they do not benefit from reduced journey-to-work costs by living there, they must find compensation in some other form to maintain utility spatially invariant. It's easy to see this from the indirect utility function which does not have spatially differential leisure available to compensate for spatially differential housing prices: $v(k) = v[y(k), p_z, p_h(k)]$. Consequently, residents will have to receive some compensation through their wage income according to where they live. Notice that, correspondingly, there are no transportation costs in the budget constraint; surely people travel throughout the city for various types of purposes, but we assume that those trips have no particular, dominant direction and consequently would give no particular spatial gradient to housing prices.

The first-order conditions for housing, the composite commodity, and the Lagrange multiplier on the budget constraint (not shown here) are the same as they were in the monocentric city model. However, the first-order condition on location is $q \bullet \Delta p_h(k) / \Delta k = \Delta y(k) / \Delta k$. We can rearrange this spatial equilibrium condition a bit to facilitate comparison of the steepness of the spatial wage gradient relative to the housing price gradient: $\% \Delta w(k) / \Delta k = [p_h(k) q(k) / y(k)] \% \Delta p_h(k) / \Delta k$,

which says that the slope of the wage gradient is a fraction of the slope of the housing price gradient, the fraction being the share of housing expenditures in wage income, which may be about $1/4$ to $1/3$ in contemporary industrial societies, possibly higher in ancient societies.

12.4.4 *Endogenous centers*

All the resource allocation that goes on in the monocentric city model revolves around the exogenously specified central business district or central transportation facility through which all exports from and imports to the city move. The version of the monocentric city model we presented in subsection 12.4.1 represented all the city's employment at a single point; other versions of the model develop price and density gradients of a land-using business sector in the interior ring of the city but still rely on a central transport node to orient the location of production facilities. This subsection tries to offer some modest reassurance that this convenient assumption about the existence of a fixed, nodal attraction could, with some effort, be generated by the same kinds of forces that already exist within the spatial model of the production facility developed in the previous chapter and modified with the concept of various external economies in subsection 12.2.1. Those models become intricate quickly so we'll give just the barest, hand-waving outline of the kinds of forces that can produce city centers endogenously.

Rather than assume that firms must ship their products to a predetermined transport node at the center of the city's business district to export them (and similarly for their imports), we can assume that firms can ship their output from their plant sites, wherever they may be located within the city, at no appreciably greater cost than is available from a central node. Next, let firms interact with one another. These interactions are more beneficial the closer firms are to one another; that is, these interaction benefits decay with distance. However, the interactions among firms occur via land-using streets which compete with buildings for urban land; that is, the interactions require land for streets. Additionally, workers must commute to firms, and firms must pay higher wages the further away their workers live.

A greater concentration of firms in one area leads to greater congestion costs and greater competition for land, both among the facilities of different firms and between buildings and streets. Similarly with attracting a labor force from greater distances. If commuting costs are low relative to the benefits firms receive from interactions with one another, a monocentric city will emerge. Depending on the type of production in which different categories of firms engage (for example, export versus local service or retail), several centers or zones of mixed residential and business locations may emerge. These are more realistic patterns of activity in the sense that they mirror contemporary observations and rely for their emergence in models on forces that are commonly considered to exist (and for which separate, empirical evidence sometimes has been found).

12.4.5 *Density gradients and the ancient city*

The populations of ancient cities have been an elusive subject. Archaeologists have recognized that a number of dwelling units does not translate simply into a recognizable number of dwellers. First, some dwelling units may have had multiple stories that are lost, and second, different units may have had different numbers of occupants. The urban spatial model offers some predictions that may be useful in some circumstances to probe archaeological remains of structures further.

We know the qualitative relationships between several density gradients. The population density gradient is the inverse of the gradient of housing-unit size. The structural density gradient should parallel the population density gradient. We will not obtain a population density gradient for any ancient city by any direct means. However, bits of the structural density gradient may remain visible and interpretable. We offer a suggestion for implementing the concept. Locate what appears to be a major structure near the center of an ancient city. Dig a transect away from that structure in either direction, all the way across the city (assuming away mundane but important contemporary issues such as acquiring enough land); if city walls interrupt its progress, continue the transect outside the walls. In the trench of this transect, look for staircases, the structural

strength of building walls, and the apparent sizes of buildings and rooms within them. (I recognize that entire buildings will not be visible from a single trench unless the trench is wide enough probably not to be called a trench; expand the trench as necessary at particular points along the transect.) Plot the progression away from the city center of building foundation strength as an indicator of the likelihood of multiple stories. Similarly plot the building size and room size within buildings, attempting to classify buildings according to probable dwelling or otherwise. See whether the function of the building – dwelling or otherwise – appears to have a distinct spatial trend. Another transect trench perpendicular to the first through the city center would be useful, resources permitting, as would further pairs of perpendicular, radial transects.

If this procedure is successful in identifying the spatial pattern of structural density, room size, or both, we still do not have a population density gradient. It may never be possible to feel certain that the population density gradient is known from the structural density gradient, but we will have reduced the problem to one of reasoning how the number of people per room, per dwelling unit, or per square foot thereof might have changed over space. We could infer a possible housing price gradient from the structural density gradient, then look for evidence on expenditure patterns that might narrow down the range of expenditures on floor-space purchases or rentals per family or per person. Some textual evidence may be available, but caution should be maintained in using relative price information from other locations and times.

12.4.6 *Wage differentials across cities*

Wages will tend to be higher in larger cities because nontradable / nontraded goods and services produced in larger cities will be more costly, and because urban consumers have to consume these goods where they are produced, it will cost more to deliver a given level of utility.¹³ We have seen that the open city model requires equalization of utility, or we could say the real wage, across cities. The higher wages in the larger cities accomplish this equalization of utilities in the face of systematically differential consumption costs.

The way this mechanism works can be seen with a simple model using a decomposition of the price of a nontraded good (or service) and the wage rate (the model is adapted from Tolley 1974, 325–327). The price of a nontraded good can be decomposed into two cost components, assuming that such goods are produced with only labor and nontraded inputs (that is, abstracting from traded inputs, the costs of which should be largely equalized across cities, within the limits of transport costs): $p_N = a_L w + F$, where p_N is the price (= production cost) of a nontraded good, a_L is the labor input per unit of this good, and F is the payments to other nontraded inputs. One major nontraded good consumed in all cities is housing. Because housing can be located in a number of sites around a city, on relatively more expensive land closer to the city center or near a port area in a noncircular city, the actual price (or rental) paid for housing can vary, but because of the way land is priced, the variations around the city will tend to be largely matched by compensating variations in access costs, which will give the package of housing and access roughly the same total price anywhere around the city. To compare how this price-cost relationship varies among cities, we can express it in terms of percentage differences: $\dot{p}_N = \alpha_L \dot{w} + \dot{f}$, where $\alpha_L = a_L w / p_N$ is the cost share of labor in the production of the good.

If the cost of changing location among cities, at least among some people, is small relative to the present value of wage differences, labor movements will leave a workday commanding the same (utility-equivalent) consumption basket across cities, leaving intercity wage differentials equal to differences in the costs of the individual components of consumption: $\dot{w} = e_T \dot{p}_T + e_N \dot{p}_N$, in which e_T and e_N are expenditure weights. This expression says that wage differentials reflect differences in a cost-of-living index. To see the effect of the production cost differentials that underlie differences in the prices of nontraded goods across cities, the previous expression for the intercity nontraded good price differential can be substituted into the expression for the intercity wage differential and rearranged: $\dot{w} = [1/(1 - s_L e_N)](e_T \dot{p}_T + e_N \dot{f})$, in which $1/(1 - s_L e_N)$ is a multiplier which expands the effect of nontraded goods prices on the wage in a city. The multiplier is larger

as both labor's cost share in the production of nontradables, s_L , and the share of nontradables in consumption, e_N , increase. The multiplier is unambiguously positive, so the sign of the wage differential, $\dot{w}$, depends on the sign of the consumption-basket term in parentheses. The difference between cities in the price of tradable goods, $\dot{p}_T$, will be somewhere between small and idiosyncratic between pairs of cities and zero as trade eliminates their price differentials to within transportation costs, so the sign of $\dot{w}$ will depend on the sign of $\dot{f}$, the cost differential of nontradable inputs into the nontradable goods. A city with higher prices (costs) of nontradable inputs – the housing-plus-access good bulking large among them – will have a higher wage, without that higher wage implying that workers in that city are better off than workers in other, generally smaller cities. This simple model could be expanded to account specifically for nontraded goods that use other nontraded goods as inputs, increasing the multiplier price effect of nontraded goods prices on wages.

12.5 Systems of Cities

The term “system of cities” may prompt many students of the city to think of a hierarchical organization such as fits naturally with the concepts of central place theory. In central place theory a city of any given order will “serve” some multiple of cities or towns of the next higher order, with “serve” interpreted quite concretely in the form of goods being distributed from larger to smaller central place. Unfortunately for the usefulness of central place theory, that model pays far less attention to where the goods distributed from large place to small place come from in the first place. As it has been formulated, central place theory doesn't incorporate any “adding-up” constraints such as “production equals consumption” or “demand equals supply” across the entire hierarchy of places. Correspondingly, the rank-size rule for cities that has emerged from central-place thinking, in which the product of a city's rank (larger cities having lower rank numbers) and its size (population) is roughly a constant, does not incorporate a role for prices to adjust diverse people's actions into aggregate consistency.¹⁴

We offer a different approach to a system of cities that provides individual behavioral underpinnings for the aggregate pattern as well as adding-up conditions that essentially distribute the entire population of a country across settlements that range from individual farms to the largest city. The model we present is that of Vernon Henderson, developed in a series of studies over nearly two decades (Henderson 1972, 1974, 1982, 1987, 1988).¹⁵ While the Henderson model possesses the essential simplicity desirable in a model, any model that begins with the behavior of individuals and ultimately accounts for aggregate patterns in a nation will have a lot of steps in it. As the reader is aware by now, this means a number of shorthand expressions for both behavioral and accounting relationships. Each of these relationships, and the types of manipulations executed with them, are familiar to the reader by now, but to assist with orientation, I offer a brief road map to the remainder of this section.

The Henderson city-system model begins by developing a model of an individual city in terms of production, consumption, and the supply of urban infrastructure. Actually, Henderson's city is a particular "type" of city, where a city's type is defined by the tradable good it produces for export. The type of city modeled as the example of the individual city is quite general, in the sense that we could study different "types" of city by seeing how different values of production parameters influence various characteristics such as population. We present this building block in subsection 12.5.1 and show how to solve for production and city size in terms of a number of technological and preference parameters. Ultimately, the characteristics of individual cities, and of the entire system of cities, reflect the technology and the culture of the society in which they reside; that's what we mean by city size and characteristics being "determined" by these parameters. We'll get more precise about those parameters forthwith. Once we get the solutions (that is, formulas for city size and other phenomena) to the individual-city model, we will examine how those equilibrium quantities are influenced by changes in those parameters. In subsection 12.5.2, we take the concept of parameter changes for a particular city and expand it to form a comparison of cities of different types – that is, cities with different characteristics. We will examine how those cities

differ in terms of their exports, costs of living and resource compositions. Subsection 12.5.3 moves up to the level of the nation or country and shows how total national production is allocated across the system of cities. Finally, we examine how the entire system of cities (the number of cities by type and size) responds to various changes, including the policies of governments that may have no intention of influencing city characteristics.

The reader may be wondering by now why this array of cities is characterized as a "system," which term implies some relationships among objects. The simplest answer is, "Because they all trade with one another." The cities produce different goods, but all the cities consume (or can consume) all of the products; the quantity of any other city's goods any city can consume depends on how much of its own goods those other cities are willing and able to consume. This is one of the adding-up constraints to which we alluded.

12.5.1 *Production and consumption within any city*

We'll employ a number of the concepts we've introduced in earlier sections of this chapter as well as in previous chapters: traded versus nontraded goods; the reminder that "specialization" in the production of a particular traded good refers to only the 40% to 50% of the labor force not engaged in production of nontradables; localization economies external to the individual firm; real versus nominal wages attributable to cost-of-living differences among locations. Each city specializes in the production of one traded good. The nontraded good we call "housing," although in any actual city other nontraded activities would go on. It may go without saying, but we need not let it, that the city is an "open city": it faces an exogenously determined level of utility because we're going to be moving a given national population around among cities.

Firms produce the traded good (the export good) according to constant returns to scale, with increasing returns at the industry level, which lets us represent the city's export production with a single, industry production function, for which we use the Cobb–Douglas form: $X = g(N)AN_0^\alpha K_0^{1-\alpha}$, in which $g(N)$ is the shift factor for the industry-level localization economies;

$\Delta g/\Delta N > 0$ and $\Delta[\Delta g/\Delta N]/\Delta N < 0$.¹⁶ N_0 is the labor time allocated to tradables production, and N in the shift function is the city's population, who divide their time between producing tradables and commuting, to be described below. K_0 is capital used in the production of the tradable good. The output elasticity α represents the degree of capital intensity of the production process, a smaller α representing a greater capital intensity.

The nontraded good, which we call housing, is produced directly with only sites (land; locations) and capital, the sites themselves being produced, as we will see, with time devoted to traveling throughout the city. A constant-returns-to-scale Cobb–Douglas production function represents the housing industry: $H = BL^\beta K_1^{1-\beta}$, in which L is sites and K_1 is capital. Land sites are produced with travel time N_2 , subject to a set of shift factors representing spatial impacts of an increasing overall population and the facilitating influence of public infrastructure capital, K_2 : $L = (DN^\delta K_2^\gamma)N_2$. As city population, N , increases, the average resident must travel further to work and in other trips, as represented by the exponential parameter δ .¹⁷ If both travel time N_2 and population doubled, sites would less than double in size because in a world with space, the additional population located at the edge of the city would have to spend more time on average traveling in an explicitly monocentric model, leaving the production of sites less efficient than with a smaller population. K_2 is public capital in infrastructure facilities that reduce the travel time required to produce sites – streets primarily, but water and sanitary facilities could have parallel effects. K_2 is considered a shift factor rather than an “instrument variable” in the production function because it is a largely irreversible investment with a large element of “once-and-for-all” about it rather than a choice variable in each time period. Overall, there are decreasing returns to site production, with $\gamma - \delta > 0$, representing a congestion effect in travel-produced sites.

Now we need to introduce some of the other variables that haven't appeared in the production functions but do enter into individual producers' profit-maximization decisions. The local wage rate is w , and it is determined within the model; p_h is the local price of housing, and p_L is the local price of sites, both of which are determined endogenously also. Exogenous to the individual

city are p_X , the national (or even international) price of the export good; and i , the national capital rental rate. Individual producers, in their profit maximization efforts, will employ the various factors of production in quantities that equalize their values of marginal product to these factor prices, given the price of the traded good. We can learn a bit about what influences various characteristics of this city by examining the relationships among these prices implied by the technical production conditions and the profit maximization efforts. We do this by showing the cost functions that correspond to the production functions.¹⁸ The cost function for the export good (under competitive conditions, its unit price) is $p_X = c_0 w^\alpha i^{1-\alpha} g(N)^{-1}$, where $c_0 = A^{-1} \alpha^{-\alpha} (1 - \alpha)^{\alpha-1}$. Remember that from the perspective of the individual city, p_X is exogenous, as is i . This exogenous export good price will be higher as capital rentals are higher, and this effect will be greater the more capital-intensive is the particular good (α is smaller with greater capital intensity, giving a larger exponential coefficient $1 - \alpha$ on the capital rental). Next, notice that the greater the scale economy effect, the lower is p_X (the exponent -1 on $g(N)$ tells us that we're dividing by $g(N)$). We can't interpret the influence of the local wage rate on the export good price in the same fashion because it is determined by this set of relationships. Invert (solve) the cost function for p_X to obtain an expression for w in terms of parameters and exogenous variables: $w = c_0^{-1/\alpha} q^{1/\alpha} i^{(\alpha-1)/\alpha} g(N)^{1/\alpha}$. With p_X and i given to the city, the wage rate rises with city size to the extent that positive scale economies prevail, and the scale effect will be greater the greater is the capital intensity of export production. This means, however, that the potential benefits to firms of the greater efficiency coming from larger industry size (the $g(N)$ factor) are dissipated through competition to hire more workers in the local labor market. Any movements in the local wage occur with a fixed level of utility, so all w represents is a nominal wage. Although variations in the nominal wage make no difference to workers, they do make a good deal of difference to firms, who face a difference in the relative price of labor and capital. The local wage will be higher the higher is the price of the export good and the lower is the capital rental (the exponent on i , $\alpha - 1$, is negative).

The cost function (unit price) for housing is $p_h = c_1 p_L^\beta t^{1-\beta}$, where $c_1 = B^{-1} \beta^{-1} (1 - \beta)^{\beta-1}$. The housing price rises as the price of sites rises and as the capital rental rises. The price of sites is also endogenous to the city: $p^L = D^{-1} \gamma^{-\gamma} w^\gamma N^\delta$. The reader can substitute the previous expression for w in the site cost function if she so desires, but as the current expression reads, a higher wage raises site rents (time costs are higher), as does a higher capital rental. Also, a larger city population raises site rents as more people compete for locations.

The production side of the city is completed by the two adding-up conditions that labor used in the export good and in site production add up to total city population and that capital used in tradables, housing, and infrastructure add up to the total capital used in the city: $N_0 + N_2 = N$ and $K_0 + K_1 + K_2 = K$. The importance of these conditions will become clearer when we move to the entire system of cities and distribute national population and capital across cities.

Let's turn to the consumption side of the city model now. Each resident has the utility function $U = E x_1^{a_1} x_2^{a_2} \dots x_n^{a_n} h^b$, in which the x_j are traded goods, one of which is produced by the type of city under analysis and exported to all the others, and h is housing. The sum of the exponential utility parameters is $\sum_{j=1}^n a_j + b \equiv f$. Individuals get their income from wage payments,¹⁹ and they pay equal amounts of tax required to pay for the public expenditures, k_2 : $y = w - iK_2/N$. To get the demand functions for the x_j and h , we take the first-order conditions of the utility maximization problem. For example, the city-wide demand for housing will be $H = byN/fp_h$. Getting the demands for each of the goods and substituting them back into the utility function yields the indirect utility function, which will be used subsequently to solve for equilibrium city size: $V = E^*(\prod_{j=1}^n p_j^{-a_j}) y^f p_h^{-b}$, in which $E^* \equiv E \sum_{j=1}^n (a/f)^{a_j}$.²⁰

The behavior of the city's government is implicit but its goal nonetheless is to maximize the utility of the residents. (The optimal quantity of public infrastructure is chosen to maximize residents' utility.) This is a simple task when we have identical residents, as we do currently. It may be thought that this objective function for the government is far from what we know of bombastic ancient governments in the Near East in particular, but let's probe this point a bit further. First, there

were several layers of government in practice, for example, in the Neo-Assyrian Empire, even though the government was unitary rather than having any formally independent components as would appear in a federal system. While governors of various cities were responsible to the central government and the emperor for revenues passed forward, they were simultaneously responsible for what passed for the wellbeing of the residents of their city, on the principle of not killing the goose that lays the golden egg. The portion of their local revenue extractions destined to preserve and enhance the wellbeing of their cities very well may have corresponded to an objective function along the lines of "maximize residents' utility," with due regard (on both our part and theirs) for different weights attached to different categories of residents and for inefficiencies in tax-farming practices. Second, even in cities under direct rule of a king or emperor without the intervening layer of a regional government, concepts of legitimacy (looking upward from the people to the ruler) and responsibility (the other direction) would have put constraints on taxation of city residents, yielding "public sector" objective functions that combined the perceived personal requirements of a royal house with concessions to the wellbeing of tax-paying residents. Third, if you still don't like this objective function, this is only a model, and other taxation and public spending principles could be implemented and examined for their consequences for the individual city and the entire city system.

This model of an individual city of a particular type can be solved for the equilibrium values of the choices its residents and government will make – that is, labor and capital allocations, total population, housing prices – in terms of production and preference parameters and exogenous variables. These solutions are found through several series of substitutions, many of which are quite intricate. We describe the methods of solution for only a few of these – the housing price (p_h), urban infrastructure capital (K_2), and total population (N). While the expressions for these solutions are relatively involved, by examining them closely, you can see how these endogenous choices will be affected by several important parameters.

From the housing supply and demand formulations we can solve for the housing price

and urban infrastructure. Equate those two expressions. Through a series of substitutions based on the marginal products of capital and labor in the housing and site production functions, and rearranging terms, we obtain an expression for the equilibrium price of housing: $p_h = c_2 p_X^{\beta/\alpha} i^{1-\beta/\alpha} N^{\delta\beta} K_2^{-\beta\gamma} g(N)^{\beta/a}$. The cost of housing, and consequently the cost of living, increases as city population increases, and decreases as more urban infrastructure facilitates movement within the city. Note also that housing costs are higher when the localization economies in the production of the export good are greater and when the price of the export good is higher. Both factors attract labor into the city, which contributes to the erosion of the productivity advantage as the cost of living rises.

Next, the equilibrium level of infrastructure is found by finding the quantity of infrastructure that maximizes utility. Recall that infrastructure (K_2) appears in the expression for after-tax income (y), which appears in the indirect utility function. Substitute this definition of income into the expression for indirect utility, as well as the formulation for w that we derived from the cost function for the export good and the solution value we just obtained for p_h above. Now, just as we proceed in taking first-order conditions, we find the value of K_2 that maximizes this expression for indirect utility; K_2 is in that resulting formulation, and we just rearrange the first-order condition to solve for it. This optimum magnitude of infrastructure is $K_2 = (\gamma\beta b/f)i^{-1}Ny$. The utility-maximizing quantity of urban infrastructure rises as population rises; as the spatial complexity parameter (γ) rises; as site intensity in housing production rises; as the preference for housing (b) is greater; and as individual, after-tax income rises. As an ordinary demand relationship, optimal infrastructure is smaller when the rental on capital is higher. The optimal choice of infrastructure is a tradeoff between increased taxes and lower housing prices; at the optimal value of K_2 , the increased tax cost is exactly counterbalanced by decreased housing prices. Figure 12.5 depicts the relationship between the volume of infrastructure and city population. The lower line identifies K_2 as increasing as N increases but as having its entire relationship with population shifted upward as either income rises or the capital rental falls.

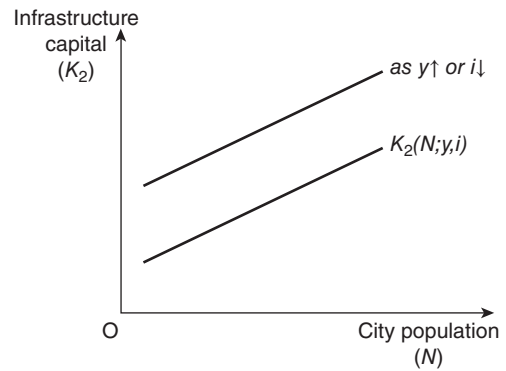


Figure 12.5 Demand for urban infrastructure. Adapted from Henderson 1988/1991, 38, Figure 2.2. © 1991 by Oxford University Press, Inc. By permission of Oxford University Press.

We can solve for the equilibrium production of the export good, X , through a series of substitutions into the original production function. Take the marginal products of capital and labor in X production and use the city full-employment conditions to find expressions for K_0 and N_0 in terms of the capital rental, the export good price, total population, and the scale economy function. An exhausting series of substitutions yields $X = c_3 g(N)^{1/\alpha} (p_X/i)^{(1-\alpha)/\alpha} N$.²¹ This tells us quite simply that export production rises linearly with city population, rises at an exponential rate somewhat in excess of 1 (assuming that $\alpha > 1 - \alpha$, which accords with virtually all evidence on capital and labor shares in output and would hold even more strongly in antiquity) as the export good price increases, and falls at the same exponential rate as the capital rental increases. Increases in the efficiency of infrastructure in producing sites (γ in the c_3 term) raise tradables production, and greater economies of scale have a strong, positive, exponential effect.

To solve for equilibrium city population, we need to express utility as a function of city size and exogenous variables and parameters. After substitution for w and y , this version of indirect utility is $V = E_1 (\prod_{j=1}^n p_j^{-\alpha_j}) [p_X g(N)/c_0 i]^{[f-b\beta(1-\gamma)]/\alpha} i^{-b} N^{\beta b(\gamma-\delta)}$, in which E_1 and c_0 are complicated terms composed of production and preference parameters.²² In this indirect utility function, remember that the p_j are the prices of export goods of other types of cities, while p_X is the export good price of the type of city we are

studying. While utility would fall if the prices of other city types' export goods rose (they are this city type's imports), utility rises as our city's export price (p_x) rises, representing an increase in income. If the rental rate on capital rises, utility falls because all input prices rise. In the same manner that we used to find the optimum value of urban infrastructure, we will take the first-order condition with respect to city population (N) that maximizes utility, but first we will offer a specific functional form for the localization economies. We let $g(N) = Ae^{-\varphi/N}$, where e is the base of the natural logarithm ($e \approx 2.71828$). This form gives a degree of increasing returns to scale that declines as N grows: $\% \Delta g / \% \Delta N = \varphi / N$, which is positive but decreases as N grows. This form of external economy lets agglomeration benefits of increasing city size eventually die out relative to the increasing consumption costs.

Now, using this form of $g(N)$ in the utility function, solving the first-order condition ($\Delta V / \Delta N = 0$) for N yields $N = (\varphi / \alpha)[f - b\beta(1 - \gamma)] / \beta b(\delta - \gamma)$. Letting the term $[f - b\beta(1 - \gamma)] / \beta b(\delta - \gamma)$ be represented by ψ , this optimal city size can be denoted $(\varphi / \alpha)\psi$, where ψ represents the internal structure of cities. This solution says that city size is a function of the internal structure of cities, the scale factor φ , and the factor intensity of export production. Note that as the scale parameter goes to zero ($\varphi \rightarrow 0$), city size goes to zero, which could be interpreted as rural production of goods with no scale effects (at least not through agglomerations of people) – basically farms and dispersed rural industry. Examining the expression for city size, we can see that if the land-intensity of housing (β) rises (that is, more land per housing unit), efficient city size falls. Similarly for increases in the spatial complexity parameter (δ), the labor-intensity of export production (α), and residents' preference for housing (b/f). If the productivity of infrastructure (γ) increases, the efficient city size rises because the cost of living will rise more slowly with population increases. And, quite intuitively, efficient city size rises the stronger are the localization economies for any size of labor force (φ). Figure 12.6 shows the inverted U-shaped relationship between utility and city size, with maximum utility reached at a city size of $N = (\varphi / \alpha)\psi$. As with the depiction of infrastructure in Figure 12.5, the lower curve shows the relationship between utility (measured

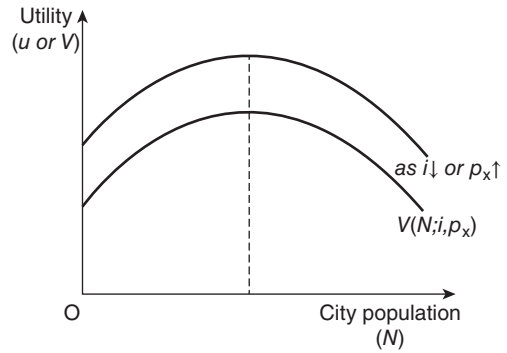


Figure 12.6 Relationship between utility and city size. Adapted from Henderson 1988/1991, 38, Figure 2.3. © 1991 by Oxford University Press, Inc. By permission of Oxford University Press.

as either U or V) and city size for given values of the capital rental and the price of the export good. The upper curve indicates that utility would be increased by either a decrease in the capital rental or an increase in the export good price.

12.5.2 Different types of cities

We have tried to remind the reader throughout the previous subsection that the individual city being modeled represents just one particular type of city in a national system of cities that contains several types of cities and several cities of each type. The model identifies a city's type by the particular export good it produces. In this subsection we examine the model to see just what differs across cities of different types and why. We index the city types with the subscript j , which refers to the export good that type of city produces, good j , which faces price p_x . A city of type j produces the quantity X_j of export good j . Each city of type j (assuming there is more than one) will have population N_j and produce the same quantity of good j , X_j . The other endogenous variables and parameters specific to a city type are also subscripted j . For instance, an important determinant of differences among city types is the labor intensity of the export good, α_j . However, the model restricts ψ , the assembly of parameters associated with consumption preferences and nontradables production, to be the same across cities.

We start with a comparison of populations of cities of different types. For any type of city, its

population is $N_j = (\varphi_j/\alpha_j)\psi$, in which φ_j is the size of the industry-level localization economies for cities producing the j^{th} traded good. The larger city size associated with larger φ_j reflects the larger city's ability to pay higher wages at any level of the capital rental. Population also increases with capital intensity of tradables production (smaller α representing a higher capital intensity) because a higher capital-labor ratio can sustain any particular wage-rental ratio with less population and consequently a lower cost of living. You can see that equilibrium city size does not depend on either overall (national) endowments of capital and labor or on factor prices; it depends strictly on parameters. Equilibrium city sizes will vary across city types, however, as φ_j/α_j varies. The largest city types are those with large-scale economies, or high capital intensities in export production, or both. Figure 12.7 shows the equilibrium populations of two types of city. The fact that each city type reaches its equilibrium size at the same, exogenous level of utility, V^* , reflects both the open-city character of the model and why cities of different sizes can coexist in equilibrium. The utility-population relationship for each city is dependent on the national capital rental and on the price of the export good each produces.

Next we look at the relationship between the prices of the export goods of different types of cities. Recall the open city assumption, which is equivalent to specifying that the country's labor market is in equilibrium: utility is the same across all city types. Then we can invert

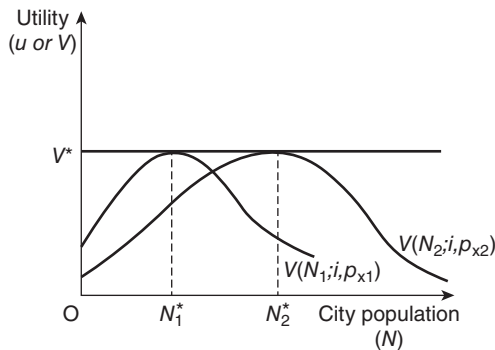


Figure 12.7 Equilibrium population of two types of city. Adapted from Henderson 1988/1991, 42, Figure 2.4. © 1991 by Oxford University Press, Inc. By permission of Oxford University Press.

the indirect utility function to solve for p_{X_j} for any good j . Do this for two types of city (which actually requires solving for only one p_{X_i} because we reference one of them with the subscript “ j ” and the other with the subscript “1,” the latter referring to the smaller city type). Divide the expression for p_{X_j} by the expression for p_{X_1} and collect the exponential terms to get $p_{X_j}/p_{X_1} = (Z_1/Z_j)^{(\alpha_1 - \alpha_j)/\alpha_1 \alpha_j}$, in which the $Z_i \equiv c_0^{-1/\alpha} g_i(N_i)^{1/\alpha} N_i^{-1/\psi}$. Examining the exponents of the capital rental rate, this expression says that an increase in i will raise the relative price of the relatively capital-intensive good. For example if X_j is more capital-intensive than X_1 (which would be reflected by $\alpha_1 - \alpha_j > 0$) and the capital rental increases, the ratio p_{X_j}/p_{X_1} rises.

Now we look for the relationship between the wage rates in different types of city. Just as we inverted the cost function for the export good to express the wage rate in terms of the unit price of the export good, we form the ratio of the wage rates in cities of types j and 1 and substitute that ratio into the ratio of export good prices above. This series of substitutions gives us $w_j/w_1 = (N_j/N_1)^{1/\psi}$. A similar series of substitutions beginning with the solution for the housing price, p_h , yields the ratio of housing prices, or the ratio of costs of living, across city types as $p_{h_j}/p_{h_1} = (N_j/N_1)^{1/\psi}$. Both wage and housing price ratios are proportional to the city size ratio, but as the numerical value of $\psi > 1$, they will rise in less than proportion to the increase in city size, as Figure 12.8 shows.²³ These two ratios have to

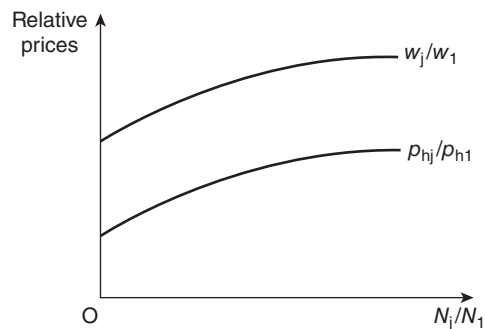


Figure 12.8 Relationship between costs and city size. Adapted from Henderson 1988/1991, 43, Figure 2.5. © 1991 by Oxford University Press, Inc. By permission of Oxford University Press.

rise in exact proportion in order to keep utility constant. This increase in the price of housing in more populous cities alerts us to the fact that residents of larger cities, even though they have the same utility as residents in smaller cities, will consume less housing, for which they compensate themselves with proportionally more tradable goods, because the relative prices of all tradable goods fall in larger cities – relative to housing. Consequently, in the archaeological remains of large and small ancient cities, we would expect to find smaller dwelling units per person or family unit in large cities than in small ones, but that would not imply that the people in the smaller cities were better off than their kinsmen in the larger cities.

The next set of comparisons deals with the overall capital intensities of cities of different types. There are several ways to measure capital intensity. So far we have referred to α , the exponential coefficient of labor in the production function for the export good, as an inverse measure of capital intensity; alternatively, $1 - \alpha$ is the corresponding, direct measure. These coefficients are not direct measures of the ratio of capital to labor (or vice versa) in that production process, as we'll show. If we let $K_0/N_0 \equiv \kappa_0$ (lower-case kappa), then the relationship between the production function coefficients and the physical capital-labor ratio in that line of production is $1 - \alpha = \kappa_0 r/w$. Thus these coefficients represent a combination of physical and price relationships. An alternative measure of capital intensity in an entire city is the ratio of capital in all uses to labor in all uses, or K/N from our full-employment conditions above. This includes the capital in housing and the labor used in travel time to produce housing sites. We can begin with the condition that the marginal product of capital in export production equals the capital rental rate, then substitute expressions for the various allocations of capital and labor into the city full-employment conditions to get an expression for the capital rental rate in terms of various parameters and the city-wide capital-labor ratio, which we designate as k : $i = p_X A(1 - \alpha)g(N)/(c_4 k)^\alpha$.²⁴ Now, equilibrium in the national capital market requires that the capital rental be equalized across cities: $i_j = i_1$ in our subscripting for different types of cities. Accordingly we equate expressions for the

capital rentals in different city types (or we can express their ratio as equal to 1) and substitute the expression we've already derived for the ratio of export prices, p_{X_j}/p_{X_1} , which gives us an expression for the ratio of k 's in different city types: $k_j/k_1 = (\Omega_1/\Omega_j)(N_j/N_1)^{1/\psi}$. From this expression, we can see that $\Delta(k_j/k_1)/\Delta\alpha_j < 0$ and, referring back to our earlier expression for the solution of city population in terms of parameters, that $\Delta(k_j/k_1)/\Delta\phi_j > 0$. In other words, the total capital intensity of cities of type j will be greater relative to the total capital intensity of cities of type 1 the greater is the technological capital intensity of the export industry in type- j cities; and similarly as the degree of scale economies in type- j cities increases. However, neither simple technological capital intensity in tradables ($\alpha_j < \alpha_1$) nor pure city size ($N_j > N_1$) by itself is adequate to assure citywide relative capital intensity ($k_j > k_1$). In other words, if we looked closely at the evidence from a pair of ancient cities and could infer relative magnitudes of their technological coefficients in tradables production, we would be unable on that basis alone to say whether the aggregate capital-labor ratio was greater in the city with the larger α .²⁵ A sufficient condition for the larger city (the type- j city) to have the relatively higher city-wide capital-labor ratio (that is, $k_j > k_1$) is that the larger city produce the relatively capital-intensive export good: $\phi_j/\alpha_j > \phi_1/\alpha_1$. What this condition implies is that, while it's possible for labor-intensive goods to be produced in the larger cities, in practice it's a lot more likely that they'll be produced in the smaller types of cities and the more capital-intensive goods will be produced in the larger city types. This may be a useful result when trying to interpret the remains of ancient city sites as a unified whole.

12.5.3 *The city size distribution and its responses to various changes*

The city size distribution is the set of ratios of numbers of each type of city. We use m_j to represent the number of cities of type j . We now know that cities of type j each have N_j population, have wages of w_j , housing prices of p_h , an export price of p_X , and an aggregate capital-labor ratio of k_j . Now we calculate how many cities there are like this, compared to any

other reference city type. To do this, we equate national supply of and demand for each type of good (two at a time). Let the demand for each traded good be $X_j^d = (a_j/f)Yp_{X_j}^{-1}$, where Y is national income after taxes. The total supply of any good X_j will be the number of cities specializing in that good times the production level per city; the ratio of the output of one good (say, good j) to that of another (good 1) then is $m_j X_j / m_1 X_1 = (a_j/a_1)(p_{X_j}/p_{X_1})$. Next, substitute the solution for the production function for any export good X_i into this expression, then substitute that result into the expression for the ratio of traded good prices to get the ratio of the number of type- j cities to the number of type-1 cities: $m_j/m_1 = (a_j/a_1)(\alpha_j/\alpha_1)(N_j/N_1)^{-1-(1/\psi)}$.

The version of the expression that contains the city populations says quite directly that the number of cities of any particular type is inversely proportional to the population of that type of city, which Figure 12.9 shows. This is satisfying in the sense that it yields some rank-size rule that corresponds to the empirical regularities of city numbers and sizes. Next, the relative number of the type- j cities (the larger ones according to our ordering scheme) to the type-1 cities is determined by the ratio of the tradables shares in consumers' budgets (the a_i terms, which come from the utility function), their ratio of labor intensities in export production (the α_i terms from the tradable goods production functions),

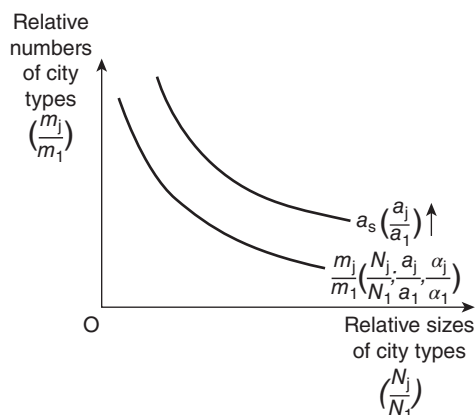


Figure 12.9 Influences on the size distribution of cities. Adapted from Henderson 1988/1991, 45, Figure 2.6. © 1991 by Oxford University Press, Inc. By permission of Oxford University Press.

and the ratio of their localization scale economies (the ϕ_i terms); the group of terms that represent the internal spatial complexity of cities (ψ) is common to all cities. The demand for good j increases relative to that for good 1 as a_j/a_1 increases, and greater demand for any good increases the number of cities that produce it. Note, however, that the demand for any good j has no direct effect on the population of the individual cities – only inasmuch as it depresses the value of b (the housing preference parameter) relative to f (the sum of the preference parameters). An increase in the ratio m_j/m_1 would entail a concentration of a country's population in its larger cities, reducing the number of smaller cities, both relatively and absolutely. What distinguishes this rank-size relationship from other models of that phenomenon is that the precise characteristics of the size distribution are attributable to specific supply and demand conditions facing the export products of the cities.

Consideration of the consequences for the city size distribution of a change in tastes among tradable goods provides some insight on mechanisms for desertion of cities and towns in antiquity as well as relative decline among well established contemporary cities. Consider a situation in which there are only two types of city, 1 and 2. When Henderson himself develops this example, he uses the case of steel-exporting cities and textile exporters, a pair of industries that could be appropriate to the ancient Near East if we substituted metals (bronze, iron, or both) for steel (Henderson 1988, 47). He lets the utility parameter a_1 rise to represent an increase in taste for metal products and a_2 fall by an equal amount ($\Delta a_1 = -\Delta a_2$, to keep $f = a_1 + a_2 + b$ and ψ constant) to represent a taste shift away from textile products.

From the solution for city population, we know that this set of changes has no effect on city size. However, it does change the composition of national output between X_1 (metals) and X_2 (textiles), the city size distribution, and relative factor prices. We must specify which industry is the relatively capital-intensive one; Henderson chooses metals, so $\alpha_2 > \alpha_1$, and for other reasons assumes that we know the type-1 city to be the larger of the two types, or $N_1 > N_2$, which guarantees that $k_1 > k_2$. We can write the full-employment conditions for the

national factor markets (as contrasted to the city-wide full-employment conditions) as the sums of total labor and capital in each type of city: $\bar{N} = m_1 N_1 + m_2 N_2$ and $\bar{K} = m_1 K_1 + m_2 K_2$, where $\bar{N}$ and $\bar{K}$ are total national labor and capital supplies. We can rearrange these two conditions to express the city size distribution ratio in terms of city and national capital-labor ratios: $m_1/m_2 = (N_2/N_1)(k_2 - \bar{k})/(\bar{k} - k_1)$, where $\bar{k} \cong \bar{K}/\bar{N}$ is the nationwide capital-labor ratio.

Now, if a_1/a_2 rises with α_1/α_2 and N_1/N_2 unchanged, m_1/m_2 has to rise according to our solution for the ratio of numbers of cities of different types. In other words, the ratio of large cities to small cities rises. The total number of cities in the economy has to fall, with no implication for total national population. Some of the textile-producing cities become metals-producing cities and grow, and the others are abandoned.

The change in relative factor prices depends on some general-equilibrium considerations such as we developed in Chapter 5. The increase in output of the relatively capital-intensive sector (metals) is accompanied by a reduction in output of the relatively labor-intensive sector. For each unit of capital released by textiles, more labor comes with it than is used by metals at its current output. Two consequences follow: metals have to get more capital from textiles for each worker it absorbs, further reducing the output of textiles, and the capital-labor ratio in metals has to fall as it expands. In fact, metals and textiles become more labor-intensive as a consequence of this demand shift. The wage-rental ratio has to fall as well to accommodate the reduction in the capital-labor ratio. The fate of utility is ambiguous when we're faced with a change in tastes to begin with, but there is an income effect from the fall in both w_1 and w_2 that tends to lower utility on that account alone. But the reduction in the wage reduces the relative price of labor-intensive textiles. This shift in taste captures much of the consequence of opening an economy to trade, which increases the demand for one of the goods by more than it does for the other.

We briefly consider several other types of exogenous change to a city size distribution. Technical changes that increased the effectiveness of infrastructure capital (γ) or decreased

the spatial drag caused by population (δ) would cause equilibrium city sizes to increase throughout the system of cities. The number of cities would have to fall. However, continued national population growth, in steady-state growth, would cause the number of cities to grow, but not the equilibrium sizes of the cities of different types.²⁶ The reader might say to herself, "Surely population growth would increase city populations!" Look once again at the form of the solution for equilibrium city size and confirm that nothing about the national capital-labor ratio enters it. Nonetheless, this result is a feature of this particular model; other formulations of this type of production-based city-size-distribution model could permit equilibrium city size to depend on capital-labor ratios.

Government policies throughout an economy will affect a country's city size distribution incidentally. As an example of a trade policy, suppose the government in the metals-and-textiles example above put an import tariff on metals, raising its domestic price. Metals production would rise and textiles production would fall, shifting population to metals cities and reducing the number of cities in the country. Additionally, the increase in the domestic price of metals would increase the relative return of the factor used intensively in it and depress that of the other factor.

Minimum wages in one form or another may have been known in the ancient world, as when a king, say, decreed that porters should receive no less than some stipulated amount, or, say, hired a large number of metal workers and paid them more than the going wage. The general equilibrium effects of such an intervention will be more far reaching than simply raising the wage of some category of worker. It can transform the national pattern of production and tilt factor rewards toward labor. Just what it does depends on the standard guiding the minimum wage. First, minimum wages tend to fix nominal rather than real wages. Second, the minimum wage may be set in terms of the average wage in a particular type of city. Depending on where in the city size distribution that city lies, cities smaller than the city that serves as the nominal wage standard will find their wages rise, and their equilibrium populations will rise to the same magnitude as the reference city's; cities larger than the reference

city are unaffected. Consequently, the number of cities must decrease, and the disappearances will come from the ranks of the formerly smaller cities. Rome's food subsidies to residents of that city would have been equivalent to wage subsidies, which, of course, raise the wage to residents of the city or cities qualifying for the subsidy. In the case of the Roman food payments, the real wages of citizen (qualifying) residents of Rome would have risen by the share of food in the typical consumer's budget – or the fraction of that budget share equivalent to the free food. This easily could have amounted to a 25–35% wage subsidy at the lower end of the income distribution, which would have included a large fraction of the city's population. This city system model tells us that the wage subsidy would have attracted additional residents to Rome to the point where the differential utility available in Rome via the wage subsidy was eroded by additional nontradable-good (housing) costs to the level of utility available elsewhere at a lower wage. Some cities at the smaller end of the city size distribution effectively would have been eliminated and their populations incorporated in that of Rome.²⁷

Subsidization of capital, either for public infrastructure or for privately employed capital, will affect city size distributions, again depending on how the subsidies are structured. The options are many. Only public infrastructure might be subsidized; all capital, infrastructure and private, might be subsidized in one specific city or one or more types of city; only private capital might be subsidized in some types of city while public infrastructure in, say, the capital city is subsidized. Subsidization of what we are calling infrastructure capital – streets and related structures that lower the cost of housing – will increase utility in the subsidized cities. However, without restrictions on population movements, migration to such a subsidized city will drive down those utility levels to what is available in the rest of the city system. If the subsidized cities are large relative to the national population, the exogenous level of utility will also fall. A uniform subsidy to infrastructure capital in all cities would have no effect on equilibrium city sizes but would distort the allocation of capital between infrastructure and other uses. A subsidy to infrastructure investments in one type of city would reduce the cost

of producing that city-type's export good, which would lead to an increase in the number of that type of city but would not affect its equilibrium population. However, after all adjustments are made, the value of this subsidy will be capitalized into wages or land rents as cities compete to join the ranks of that type of city. It does lead to an increase in urban concentration though.

Several final observations on city size distributions. First, in "finding" locations, larger cities will tend to capture sites with greater amenities as perceived by both producers and consumers, ranging from better water and more pleasant climate to access to water transportation routes. From the consumer's perspective, the services provided by the amenity raise utility, which causes city size to expand, raising housing prices so as to push utility back down to the exogenous level. Since amenities are public goods, larger cities have more people from whom payment can be exacted through higher rents. From the producer's perspective, firms find they are able to bid more for labor because they can produce more for a given level of inputs than can firms in cities less well endowed with amenities, but doing so drives up their wage costs as city size expands, and their marginal cost advantage is dissipated. Nonetheless, sites with greater levels of valued amenities will be occupied before sites with lower levels.²⁸

Second, city systems in politically centralized societies will tend to be more concentrated – that is, have more of their populations bunched up in a smaller number of larger cities, possibly even in a single, "primate" city – than less centralized societies. Several mechanisms contribute to this result. First, the more centralized political systems tend to have less well developed (and fewer) mechanisms for identifying the public service demands of populations outside the national capital region. Second, related to the first mechanism, people making the decisions about the national distribution of public infrastructure may be able to insulate themselves personally from the ill effects of increasing city size; additionally they are likely to be land owners and thence benefit from increasing rents. Third, what we today would be inclined to call corruption may be less constrained by competition in more centralized governments. Officials of the central government may be able to place themselves

in critical positions in the provision of public services and infrastructure construction and operation in cities elsewhere in the country, diverting resources from actual service provision and thus depressing the sizes of the dispersed cities and towns. Finally, notions of hierarchy borrowed from government organization may be applied inappropriately to thinking about the effectiveness of locations of various public investments, relying on “trickle-down” mechanisms to deliver benefits to locations outside the capital region. Less political centralization puts more constraints on the ability of bureaucrats to satisfy their personal tastes on these matters.

Third, when countries are open to trade abroad, the income base forming the demand for individual cities’ tradable products exceeds the income of any one of the nations alone. One of the implications of this fact is that the proper geographical boundary for a system of cities may encompass more than a single country. This aspect of city size distributions has not been explored to my knowledge.

Finally, the cities that comprise the system of cities in Henderson’s model must be interpreted liberally to incorporate the functional types of cities identified in the Noyelle–Stanback taxonomy of contemporary U.S. cities. Services comprise a substantial portion of the employment of many cities. The purely local services we can associate with the nontradable good, which is called “housing” in the Henderson model, demonstrate some of the important, spatial characteristics of housing. Some cities export services. While very recent developments in communications technology have expanded the scope for service exports, some services have been exported for centuries: financial services and shipping and other transportation services, education, and government administration. Other cities, particularly toward the small end of the size distribution, characterized as regional mixed service centers, provide a range of manufactured and service exports to surrounding agricultural regions rather than an entire country.²⁹ Finally, we may question the empirical correspondence between the model’s prediction that larger cities will have more capital-intensive export industries and higher overall capital-labor ratios than smaller cities. Some of the largest cities in the contemporary world are quite diversified while the mid-size

and smaller cities have greater concentrations of their employment in one or a few industries. Henderson himself has taken the distinction between production processes tied to resource bases and others that are footloose and with it developed a model of metropolitan areas, consisting of a core city, typically attracted to some resource base, and a group of spatially contiguous, smaller cities that derive transportation-cost savings from proximity to the markets of the large, capital-intensive resource-based cities (Henderson 1988, 190–193). This type of multi-city agglomeration certainly characterizes the world’s contemporary urban system, and the categories in terms of which it leads us to think may be useful in adapting his city-system model to ancient city systems.

12.6 Urban Finance

When archaeologists survey the monumental remains of ancient cities – the cyclopean walls of the Mycenaean Greek cities, the temples of the Mesopotamian and Egyptian cities – they tend to say, “What an ingenious, industrious people!” Some may add, “What a surplus they had!” An economist looking at the same remains would be inclined to say, “How in the world did they *pay* for it?!” By saying this, the economist isn’t expressing doubt or even amazement that the society had the resources to construct the monuments, as the archaeologist quoted second well recognizes. What she’s pondering is how various elements of the society managed to pull off the decisions to allocate such sizeable concentrations of resources to these structures. These structures imply the existence of allocation mechanisms capable of coordinating diverse components of society as well as the decision to allocate some proportion of potential private consumption to “public” consumption.³⁰

This section addresses in abbreviated fashion some topics in the public finance of cities that have been considered important in contemporary times. The analytical devices that have been developed to study some of these topics are tailored nearly skin-tight to contemporary institutional structures, but we will try to distill from these models what may be useful for thinking about ancient cities with different integumentary

systems if not necessarily different skeletal systems. The subjects are what services ancient cities are likely to have provided; how the basic forces of scarcity would have acted upon the supply of and demand for these items; and what the consequences of raising the revenue to pay for these items would have been, given the technological options of the times.

Before embarking, however, it is useful to offer a quantitative dimension based on contemporary magnitudes. In the United States in 1981–1982, about 12% of personal income went to local (city, excluding state) taxes, which in turn went to city expenditures. About 40% of that went to education, which was not publicly provided until quite recently but has, of course, decreased since the early 1980s. The categories of contemporary city expenditures that are likely to have had some correspondence in ancient cities include: public safety, 8%; welfare (feeding the hungry), 4.7%; health and hospitals, 6.9%; parks and natural resources, 2.4%; sanitation (drinking water and waste disposal), 4.7%; transportation (street maintenance), 4.7%; and administration and interest, 8% (Mills and Hamilton 1989, 295, 308, Tables 13.3, 13.6). This subset of urban expenditures would amount to some 2.9% of contemporary personal income. It's well known, of course, that many ancient cities had rudimentary waste disposal systems and services, and some of what may have existed was likely privately provided. The allocation of treatment of the sick between private and public expense is an open question and probably subject to variation across the expanses of time and place. Developed open spaces existed in ancient cities, and even if it's difficult to consider them police forces, labor was allocated to various aspects of public order and security (Green and Teeter 2013, 38–41). Administration is just government, and “interest” can be translated into the amortization of the payments for the city walls, ziggurats, megara, palaces, aqueducts, and other public structures.³¹ Between reductions of this 2.9% figure for what the ancient cities did not or may not have provided that contemporary cities do and additions to account for city walls, higher fixed costs with a smaller proportion of the total population living in cities, and lesser technological ability (that is, higher relative cost) to produce some of the items that were provided, around 3% may not be

an outrageous figure to think of as the share of personal income of urban residents on which this section focuses.

12.6.1 *Local public goods*

As we introduced in Chapter 6, local public goods are a category of public goods which, in their extreme, or “pure,” form, are inexhaustible and nonexcludable: one person's consumption doesn't subtract from another's consumption; and, once provided for anyone, excluding any person's consumption is difficult or impossible. National defense is the prototypical, pure public good on these criteria. Local public goods share these characteristics but not in a pure form; and exclusion can be accomplished by the geography of provision – that is, if you don't live in the right place, you can't consume it, or at least not without paying a fee (Tiebout 1956).³² As is the case with the pure public good, the optimal supply is determined by the sum of individuals' willingness-to-pay. In other words, we add individuals' demand curves vertically instead of horizontally; again, see Chapter 6.

Let's take some examples of local public goods provided by the ancient city. One that is virtually without dispute when evidence can be found is the city fortification wall. This supplied the city equivalent of national defense, and undoubtedly for many residents living outside the citadel circuit wall as well as those dwelling within. Maintaining and manning (and womaning too when the occasion rose) the walls were also local public goods.³³ In a city that was part of a larger state, many laws would have been of no more than local interest, such as restrictions on building characteristics, designated clear spaces near the city gates, rules for location or occasional participation in the agora or bazaar, and so on. Establishing, actively enforcing, and being available to adjudicate on enforcement all are costly activities. City protective actions in time of disease or famine also would be costly (not in the sense of especially expensive in comparison to something else, like silk purses, but simply requiring the diversion of real resources away from other activities).

Aqueducts, both in Rome and the Near East, probably qualify as local public goods even though their construction may have been

conducted and financed by “national” rather than strictly municipal officials. National officials acting on behalf of a municipality, particularly when legal distinctions between levels of government were not especially clearly delineated, are in analytical terms supplying local public goods in these cases. The well known, large-scale refuse removal from Rome exemplifies public sanitation efforts.

Enforcement of weights and measures may have been a “national” responsibility. If the Neo-Assyrian town councils are any guide to other circumstances and practices, city enforcement of accepted commercial practices, and of business interests in general, may have involved the provision of some regularly established agencies that, for our analytical purposes, would qualify as “municipal agencies.” Certainly Athens supervised and regulated weights and measures, as noted in Chapter 6.

While contemporary public libraries are examples of local public goods, the Libraries of Alexandria and Pergamum surely were more than local public goods, more along the lines of the U.S. Library of Congress or the British Library than a local public lending library. Some of the apparent libraries discovered in tablets in ancient Near Eastern sites may have been private. However, even the private libraries may have been examples of privately supplied club goods based on some kind of membership other than simple geographical residence. Schools at this time were, by and large, private (Harris 1989, 17).

In the cases of city states, such as Athens and Corinth in Classical Greece and Late Bronze Age Ugarit on the Levantine coast, where the actual, eponymous cities stood distinct from relatively small hinterlands, the term “local public good” is best applied to the city alone. The road system, for instance, throughout the Corinthia would have been a Corinthian public good but not a local public good of the city of Corinth. The city walls of Athens would have been a local public good of the city of Athens but not a public good conferring benefits on all Athenians regardless of their residence although many did squeeze in during the Peloponnesian Wars.

We have identified several types of local public goods not as a prelude to some exhaustive catalogue or detailed analysis of the optimal provision of any one of them but as examples

to demonstrate that the concept of the local public good has real, interesting content in the case of the cities of the ancient Mediterranean. Correspondingly, the analysis of the allocation mechanisms by which they were provided, as well as the economic consequences of their provision, are topics of genuine interest.

12.6.2 *What to supply and how much*

This subsection focuses on how associative groups like cities and towns decide on what to provide themselves as local public goods and how they determine how much of each to provide. In contemporary urban economic theory, there are two primary models of this process, the Tiebout model of “voting with one’s feet,” and the median voter model of voting with the ballot.³⁴ Both are suited primarily to democratic forms of government and jurisdictional fragmentation such as the contemporary Western system of central cities and independent suburbs, but both contain interesting mechanisms that have more general applicability and lessons.

The Tiebout model was developed as a mechanism for solving the preference-revelation problem for public goods that Paul Samuelson had demonstrated in a now-famous paper the previous year: while we know that the optimal provision of a public good is the quantity that equates the supply curve of the public good with the vertical sum of individual demands for it, how can we ever get those individuals to tell honestly what their demands (willingness to pay) are when each can benefit by understating his or her own preferences (assuming, of course, that everybody else is honest!)?³⁵ Tiebout’s model took the public-good provision problem to the U.S. metropolitan region of the mid-1950s and later, with many suburbs politically independent of one another and of the central city they surround. Each of these suburbs provides an array of public goods. Individuals (families) differing in income and tastes³⁶ face a large number of suburbs that provide different arrays of public goods. In any suburb, the individual can purchase a combination of housing and public goods at the going price of housing and local tax on housing, which is sufficient to just pay for the public goods. Their choices of locations reveal their demands for local public goods, and the

competition among suburbs for residents keeps the local public goods providers efficient. Zoning restrictions that put a floor on minimum house or lot sizes will keep lower income people from locating in suburbs with higher income residents and free-riding on the local public goods with an inexpensive house, the tax on which fails to cover the cost of the public good consumption enjoyed. The subsequent 40 years of empirical scrutiny and theoretical development of the Tiebout model leave it as an interesting mechanism that is unlikely to be an efficient provider of local public goods even if some of its assumptions are not always met. It does seem to characterize an important component of the provision of local public goods and services.

One of the more important objections to one of the Tiebout model's assumptions focuses on the mobility required for residents to be able to move among suburbs. At the other extreme from perfect mobility, the median voter model lets housing consumers with divergent preferences for public goods be perfectly immobile but able to vote for officials who will direct compliant bureaucrats to select a designated quantity or array of public goods. The model derives its name from the fact that local expenditure and taxation decisions replicate the preferences of the "median voter," the person whose opinions are in the very middle of the diversity of opinion on these matters among her neighbors. Again, further examination of this model has offered modifications, among them the ability of nonmedian opinions to have some influence on decisions, although the opinions of people far from the median drift into insubstantiality.

Despite the differences in the political circumstances of the Mediterranean and the Aegean regions 2000 to 4000 years ago, there are reasons for keeping both these models in mind when thinking about how decisions regarding local public good provision was accomplished in those times and places. Both models remind us that any agent (king, emperor, governor, high priest) in a position to make decisions about local public good provision in the ancient Mediterranean world could have dissatisfied his or her subjects to some degree with the choices. On the principle that more satisfied subjects are more malleable than less satisfied ones, there are reasonable grounds for even an absolute sovereign to think

at least about the tradeoffs between his own current consumption and the wellbeing, possibly even the happiness, of his subjects. To the extent that some of the local public goods – possibly most of the ones that are likely to have been provided in antiquity – also made the king's revenue sources more productive, there would have been further grounds for soliciting opinions, even if only those of the wealthy lords.

On the spatial specifics of the Tiebout mechanism, not every resident needs to be mobile for the mechanism to work. The marginal consumers – the ones who are close to their tipping points about choosing one location over another – are the ones who affect land values in alternative locations. Attraction of immigrants at the margin is in the interests of whoever benefits from having more people, and possibly more productive people, in a city. If the governor of a city that is one of several dozen in a country or kingdom or empire either owns land in the city he governs or derives income, wealth, or consumption in proportion to the value of production in the city, he would leave a free lunch on the table by ignoring factors under his control that would make location in his city more attractive relative to other cities.

12.6.3 *Raising revenue*

Most city revenue comes from taxes, user fees, and outside grants. The dominant contemporary tax revenue source is the property tax: annual taxes on housing units, other buildings, and possibly vacant private land. These three sources are likely to have been the principal revenue sources in antiquity as well.³⁷ User fees could have ranged from a small charge per donkey or porter entering one of the city gates to "licenses" for vendors in the public market. For outside grants, the royal treasury would be a logical source to permit the supply of goods deemed desirable for a city that proved beyond its individual revenue capacity. A central treasury might advance loans for construction of a large wharf complex in a relatively small riverine city. For the remainder of this subsection, however, we concentrate on local taxes.

Taxes on housing (on other structures as well) tend to depress property (capital) values and reduce the supply of housing. If capital is perfectly immobile between jurisdictions, the

relationship between housing value and property tax has the form $V = (pH_f - tV)/i$, where V is the stock value of a unit of housing, p is the annual rental, H_f is the flow of housing services, i is the interest rate, and t is the tax rate. Rearranging this, we get $V = pH_f/(i + t)$, in which the tax rate t depresses housing value V .

Figure 12.10 shows the effect of a tax on housing rental prices, housing stock prices, and the quantity of housing stock. We show the supply and demand for the stock of housing in the lower panel and for the flow of housing services in the upper. We can make the two sets of curves correspond. Begin by drawing the supply curve of housing stock, S_s , in the lower panel and the demand curve for flows of housing services, D_f , in the upper panel, both of which are observable. To convert units of stocks to a flow supply curve, multiply the value of V at each point on the stock supply curve, S_s , by i to get the corresponding value of p in the upper panel, denoted S_f . Next, convert the flow demand for housing into a stock demand curve by dividing each value of p on the flow demand curve D_f by i to get the

corresponding V , the points of which trace out the stock demand curve, D_s , in the lower panel. Stock and flow equilibrium housing quantities are designated by H_0 (the quantity is the same whether you rent it or buy it).

Now impose a tax on housing. Housing value becomes $V = pH/(i + t) < pH/i$. This does not change the flow demand for housing services, but it does raise the flow supply curve of housing services in the top panel. To get the new flow supply curve, multiply each point V on the stock supply curve by $i + t$ to get S_f^t . However, nothing about the underlying conditions of the supply of housing units has been affected by the tax, so the stock supply curve remains unchanged. However, the long-run demand for housing units will be affected because the equilibrium price of housing service flows has risen. To get the new stock demand curve, divide p on the flow demand curve by $i + t$ to get the new stock demand curve, D_s^t . The equilibrium rental payment for housing services rises, while the flow of housing services, the stock of housing services, and the equilibrium housing value all fall. If tenants and housing owners are different persons, tenants' rental payments go up while owners' returns fall. The power of the property tax to reduce its tax base can be seen clearly by observing that the tax affects both the price and quantity terms of that base, VH .³⁸

Consider the alternative circumstance in which capital is perfectly mobile – possibly in the long run. Figure 12.11 shows a much flatter stock supply curve of housing. The only reason the supply curve is not perfectly flat is because immobile land is an input to housing. With perfectly mobile capital, an increase in a tax rate will have a less than proportional, negative influence on the stock housing price³⁹ but a far greater negative effect on the land price.⁴⁰ However, in this case, housing owners bear a much smaller share of the tax payment than housing renters.

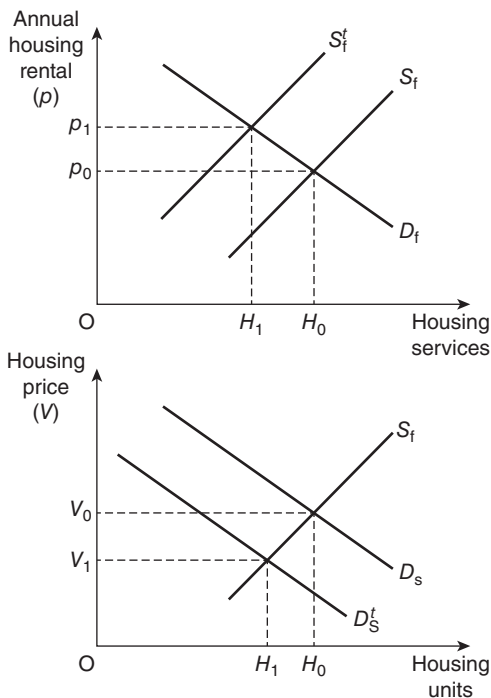


Figure 12.10 Effect of an urban property tax on rentals, prices and quantities of housing.

12.7 Suggestions for Using the Material of this Chapter

Densities should vary systematically across an ancient city. Occupation should be expected to be denser around attraction sites: public buildings (temples, administrative buildings), major

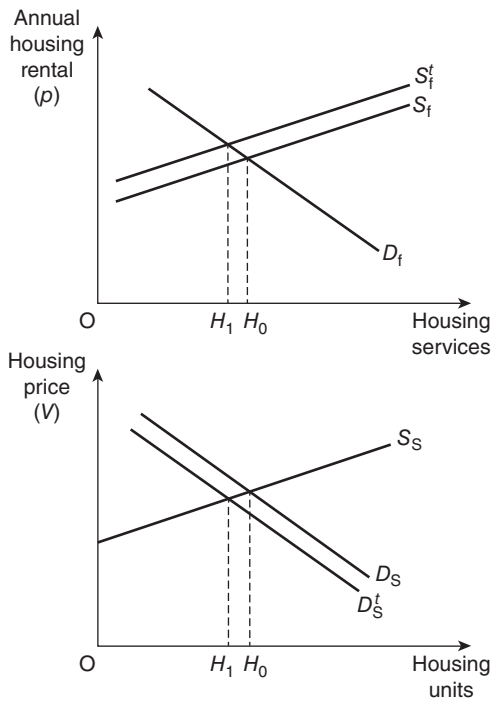


Figure 12.11 Effect of urban taxes when capital is perfectly mobile.

markets or bazaars, possibly around city gates, maybe around bridges across a river within a city.

Living space (dwelling sizes) will vary without necessarily implying wealth or income differences of the occupants: dwelling size will depend

on the size of the city in which the dwellings are located and where the dwellings are located within each city.

Clustering and location of certain activities within an ancient city may have resulted from negative externalities (smoke, smells, noise) – for example urban metallurgical establishments, dying and tanning.

Buildings were major capital structures – they lasted a long time. There were many people to shelter. There must have been considerable complexity of social organization even as far back as the Late Neolithic-Final Neolithic-Early Helladic towns such as Manika on Euboea, Knossos, and so forth. Laws (or codified custom, whatever we call it today) must have existed to govern the right to build on land, assignment of the right to use a building once built (ownership), and succession (because owners eventually died, if they didn't leave before). Archaeologists are accustomed to finding buildings within a settlement that were abandoned at different dates, occupied buildings coexisting in proximity to buildings abandoned and sometimes even used as burial sites. The cost of tearing down a building, even a post-and-beam construction with post-hole foundations, can be expensive, so why bother unless the land needs to be used? Letting an abandoned building continue to fall down near currently occupied buildings suggests a low price of land, as well as possibly a low awareness of fire hazards.

References

- Adams, Robert McC. 1981. *Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates*. Chicago IL: University of Chicago Press.
- Bang, Peter Fibiger. 2008. *The Roman Bazaar; A Comparative Study of Trade and Markets in a Tributary Empire*. Cambridge: Cambridge University Press.
- Blanton, Richard E. 1994. *Houses and Households; A Comparative Study*. New York: Plenum.
- Brueckner, Jan K. 1987. "The Structure of Urban Equilibria: A Unified Treatment of the Muth-Mills Model." In *Handbook of Regional and Urban Economics*, Vol. 1. *Urban Economics*, edited by Edwin S. Mills. Amsterdam: North-Holland, pp. 821–845.
- Carey, Christopher. 2000/2001. *Democracy in Classical Athens*. London: Duckworth.
- Cornes, Richard A. and Todd Sandler. 1986. *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge: Cambridge University Press.
- Delougaz, Pinhas, Harold D. Hill, and Seton Lloyd. 1967. *Private Houses and Graves in the Diyala Region*, OIP LXXXVIII. Chicago IL: University of Chicago Press.
- Doumas, Christos G. 1983. *Thera, Pompeii of the Ancient Aegean*. London: Thames & Hudson.
- Engels, Donald. 1990. *Roman Corinth; An Alternative Model for the Classical City*. Chicago IL: University of Chicago Press.

- Erdkamp, Paul. 2001. "Beyond the Limits of the 'Consumer City': A Model of the Urban and Rural Economy in the Roman World." *Historia* 50: 332–356.
- Erdkamp, Paul. 2008. "Mobility and Migration in Italy in the Second Century BC." In *People, Land, and Politics: Demographic Developments and the Transformation of Roman Italy 300 BC–AD 14* (Mnemosyne supplement 303), edited by Luuk de Ligt and Simon Northwood. Leiden: Brill, pp. 417–449.
- Ermisch, J.F., J. Findlay, and K. Gibb. 1996. "The Price Elasticity of Housing Demand in Britain: Issues of Sample Selection." *Journal of Housing Economics* 5: 64–86.
- Finley, M.I. 1980. *Ancient Slavery and Modern Ideology*. New York: Viking.
- Finley, M.I. 1985. *The Ancient Economy*, 2nd edn. Berkeley CA: University of California Press.
- Frankfort, H., and J.D.S. Pendlebury. 1933. *The City of Akhenaten*, Part II. *The North Suburb and the Desert Altars. The Excavations at Tell el Amarna during the Seasons 1926–1932*. London: Egypt Exploration Society and Oxford: Oxford University Press.
- Frier, Bruce W. 1977. "The Rental Market in Early Imperial Rome." *Journal of Roman Studies* 67: 27–37.
- Frier, Bruce W. 1980. *Landlords and Tenants in Imperial Rome*. Princeton NJ: Princeton University Press.
- Goldsmith, Raymond W. 1962. *The National Wealth of the United States in the Postwar Period*. Princeton NJ: Princeton University Press.
- Gomme, A.W. 1937. "Traders and Manufacturers in Greece." In *Essays in Greek History and Literature*, by A.W. Gomme. Oxford: Basil Blackwell, pp. 42–66.
- Green, Jack, and Emily Teeter. 2013. *Our Work. Modern Jobs – Ancient Origins*. Oriental Institute Museum Publications 36. Chicago IL: Oriental Institute of the University of Chicago.
- Harris, William V. 1989. *Ancient Literacy*. Cambridge MA: Harvard University Press.
- Henderson, J. Vernon. 1972. *The Types and Sizes of Cities: A General Equilibrium Model*. Ph.D. dissertation, University of Chicago, Chicago IL.
- Henderson, J. Vernon. 1974. "The Sizes and Types of Cities." *American Economic Review* 64: 640–656.
- Henderson, J. Vernon. 1982. "Systems of Cities in Closed and Open Economies." *Regional Science and Urban Economics* 12: 325–350.
- Henderson, J. Vernon. 1985. *Economic Theory and the Cities*, 2nd edn. New York: Academic Press.
- Henderson, J. Vernon. 1987. "The Analysis of Urban Concentration and Decentralization: The Case of Brazil." In *The Economics of Urbanization and Urban Policies in Developing Countries*, edited by George S. Tolley and Vinod Thomas. Washington, D.C.: World Bank, pp. 87–93.
- Henderson, J. Vernon. 1988. *Urban Development; Theory, Fact and Illusion*. New York: Oxford University Press.
- Hochman, Oded. 1977. "A Two-Factor, Three-Sector Model of an Economy with Cities: A Contribution to Urban Economics and International Trade Theories." Mimeo, Department of Economics, University of Chicago, Chicago IL.
- Hochman, Oded, and Haim Ofek. 1977. "The Value of Time in Consumption and Residential Location in an Urban Setting." *American Economic Review* 67: 996–1003.
- Hopkins, Clark. 1979. *The Discovery of Dura Europos*. New Haven: Yale University Press.
- Hopkins, Keith. 1980. "Taxes and Trade in the Roman Empire (200 B.C. – A.D. 400)." *Journal of Roman Studies* 70: 101–125.
- Jashemski, Wilhelmina F. 1979. *The Gardens of Pompeii; Herculaneum and the Villas Destroyed by Vesuvius*. New Rochelle NY: Caratzas Brothers.
- Lo Cascio, Emilio. 2003. "Did the Population of Imperial Rome Reproduce Itself?" In *Urbanism in the Preindustrial World; Cross-Cultural Approaches*, edited by Glenn R. Storey, 52–68. Tuscaloosa AL: University of Alabama Press.
- Loomis, William T. 1998. *Wages, Welfare Costs and Inflation in Classical Athens*. Ann Arbor MI: University of Michigan Press.
- McDonald, John F. 1997. *Fundamentals of Urban Economics*. Upper Saddle River NJ: Prentice-Hall.
- Meiggs, Russell. 1973. *Roman Ostia*, 2nd edn. Oxford: Clarendon.
- Migeotte, Léopold. 1997. "Le contrôle des prix dans les cités grecques." In *Entretiens d'archéologie et d'histoire, 3: Prix et formation des prix dans les économies antiques*, edited by J. Andreau, P. Briant, and R. Descat. Toulouse: St.-Bernard-de-Comminges, pp. 33–52.
- Mills, Edwin S., and Bruce W. Hamilton. 1989. *Urban Economics*, 4th edn. Glenview IL: Scott, Foresman.
- Murray, A.T. 1936. Translator, *Demosthenes IV. Private Orations XXVII – XL*. Loeb Classical Library. Cambridge MA: Harvard University Press.
- Muth, Richard F. 1969. *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*. Chicago IL: University of Chicago Press.
- Nevett, Lisa C. 1999. *House and Society in the Ancient Greek World*. Cambridge: Cambridge University Press.
- Noyelle, Thierry, and Thomas Stanback, Jr. 1983. *The Economic Transformation of American Cities*. Totowa NJ: Rowan & Allanheld.
- Owens, E.J. 1991. *The City in the Greek and Roman World*. London: Routledge.

- Paine, Richard R., and Glenn R. Storey. 2003. "Epidemics, Age at Death, and Mortality in Ancient Rome." In *Urbanism in the Preindustrial World; Cross-Cultural Approaches*, edited by Glenn R. Storey. Tuscaloosa AL: University of Alabama Press, pp. 69–85.
- Peet, T. Eric, and C. Leonard Woolley. 1923. *The City of Akhenaten*, Part I. *Excavations of 1921 and 1922 at Tell el-Amarna*. London: Egypt Exploration Society.
- Pendlebury, J.D.S. 1951. *The City of Akhenaten*, Part III. *The Central City and the Official Quarters. The Excavations at Tell el-Amarna during the Seasons 1926–1927 and 1931–1936*. London: Egypt Exploration Society and Oxford: Oxford University Press.
- Pirenne, Henri. 1925. *Medieval Cities; Their Origins and the Revival of Trade*. Translated by Frank D. Halsey. Princeton NJ: Princeton University Press.
- Pirson, Felix. 1997. "Rented Accommodation at Pompeii: The Evidence of the Insula Arriana Polliana VI 6." In *Domestic Space in the Roman World: Pompeii and Beyond*, edited by Ray Laurence and Andrew Wallace-Hadrill. (*Journal of Roman Archaeology* supplement). Portsmouth NH: Journal of Roman Archaeology, pp. 165–181.
- Pirson, Felix. 1999. *Mietwohnungen in Pompeji und Herkulaneum; Untersuchungen zur Architektur, zum Wohnen und zur Sozial- und Wirtschaftsgeschichte der Vesuvstädte*. Munich: Dr. Friedrich Pfeil.
- Postgate, J.N. 1992. *Early Mesopotamia; Society and Economy at the Dawn of History*. London: Routledge.
- Reynolds, Susan. 1977. *An Introduction to the History of English Medieval Towns*. Oxford: Clarendon Press.
- Rubinfeld, Daniel L. 1987. "The Economics of the Local Public Sector." In *Handbook of Public Economics*, Vol. 2, edited by Alan J. Auerbach and Martin Feldstein. Amsterdam: North-Holland, pp. 571–645.
- Scheidel, Walter. 2003. "Germs for Rome." In *Rome the Cosmopolis*, edited by Catharine Edwards and Greg Woolf. Cambridge: Cambridge University Press, pp. 158–176.
- Scheidel, Walter. 2004. "Human Mobility in Roman Italy, I: The Free Population." *The Journal of Roman Studies* 94: 1–26.
- Sjoberg, Gideon. 1960. *The Preindustrial City; Past and Present*. New York: Free Press.
- Stöger, Hanna. 2011. *Rethinking Ostia: A Spatial Enquiry into the Urban Society of Rome's Imperial Port-Town*. Archaeological Studies Leiden University 24. Leiden: Leiden University Press.
- Straszheim, Mahlon. 1987. "The Theory of Urban Residential Location." In *Handbook of Regional and Urban Economics*, Vol. 2. *Urban Economics*, edited by Edwin S. Mills. Amsterdam: North-Holland, pp. 717–757.
- Temin, Peter. 2006. "Estimating GDP in the Early Roman Empire." In *Innovazione tecnica e progresso economico nel mondo romano*, edited by Elio Lo Cascio. Rome: Bari, pp. 31–54.
- Tiebout, Charles M. 1956. "A Pure Theory of Local Public Expenditures." *Journal of Political Economy* 64: 416–424.
- Tolley, George S. 1974. "The Welfare Economics of City Bigness." *Journal of Urban Economics* 1: 324–345.
- Turnbull, Geoffrey K. 1995. *Urban Consumer Theory*. Washington, D.C.: Urban Institute Press.
- Wallace-Hadrill, Andrew. 1994. *Houses and Society in Pompeii and Herculaneum*. Princeton NJ: Princeton University Press.
- Wallace-Hadrill, Andrew. 2011. *Herculaneum: Past and Future*. New York: Francis Lincoln.
- Weber, Max. 1958. *The City*. Translated by Don Martingale and Gertrud Neuwirth. Glencoe IL: Free Press.
- Wheaton, William C. 1974. "A Comparative Static Analysis of Urban Spatial Structure." *Journal of Economic Theory* 9: 223–237.
- Wheaton, William C. 1979. "Monocentric Models of Urban Land Use: Contributions and Criticisms." In *Current Issues in Urban Economics*, edited by Peter Mieszkowski and Mahlon Straszheim. Baltimore MD: Johns Hopkins University Press, pp. 107–129.
- Wieand, Kenneth F. 1987. "An Extension of the Monocentric Urban Spatial Equilibrium Model to a Multicenter Setting: The Case of the Two-Center City." *Journal of Urban Economics* 21: 259–271.
- Wolpert, Andrew, and Konstantinos Kapparis. 2011. *Legal Speeches of Democratic Athens: Sources for Athenian History*. Indianapolis IN: Hackett.
- Zanker, Paul. 1998. *Pompeii: Public and Private Life*. Translated by Deborah Lucas Schneider. Cambridge MA.: Harvard University Press.
- Zodrow, George R., ed. 1983. *Local Provision of Public Services: The Tiebout Model after Twenty-Five Years*. New York: Academic Press.

Suggested Readings

- Henderson, J. Vernon. 1985. *Economic Theory and the Cities*, 2nd edn. New York: Academic Press.
- O'Sullivan, Arthur. 2000. *Urban Economics*, 4th edn. Burr Ridge IL: Irwin McGraw-Hill.
- Segal, David, 1977. *Urban Economics*. Homewood IL: Irwin.

Notes

- 1 Henri Pirenne (1925), published from lectures delivered at several American universities in the fall of 1922. Weber (1958, 68–69), first published in *Archiv für Sozialwissenschaft und Sozialpolitik* 47 (1921), appears to have had in mind an export-base concept whereby rentiers' capital exports (receipts of income from outside the city) form the base for, but not the whole of, city income, the alternative interpretation being one of derived demand, in which rentiers' income formed the whole of city income, portions of which served as payments to people serving them; see the brief treatment of export base theory in subsection 12.2.3. See Finley (1985, 124–139, esp. 132) for Finley's misplaced criticism of Gomme's statement that the fifth-century Athenians were aware of the adding-up constraints involved in trade: Gomme (1937, 45) and Engels (1990). Many of the suppositions about cities and economic activities in them that Pirenne and Weber believed to be facts in the first two decades of the last century have been found to differ from more recent evidence; for a useful, if now somewhat dated, overview, see Reynolds (1977). Sjöberg (1960, Chapter 7) offers useful details about various aspects of economic activity in cities outside the industrialized West but does not address the basic accounting questions at issue here. Owens (1991, 3) contends that, "The importance of economic activity in the ancient city is often understated. . . . Ultimately the physical development of the city was conditioned by its resources." His subsequent discussion of the physical remains of a number of Greek cities, from the Geometric through the Classical periods, notes in passing direct and circumstantial evidence of production. Meiggs (1973, 270–278) characterizes the evidence of diverse, small-scale industry as well as large-scale baking and warehousing activities in Ostia, the port city of Rome, dating from the second century B.C.E. through the fourth century C.E.
- 2 Ancient historians have difficulty leaving the taxonomic consumer city model behind as they move into the twenty-first century. In a recent defense that denies being a defense of Finley's clamping onto Weber's consumer city model as the type of city antiquity generated, Bang (2008, 47–48, nn. 88 and 89) refers to Erdkamp's (2001) effort to relate that model to Hopkins's (1980) precocious analysis of interregional capital flows. Erdkamp struggles mightily with accounting concepts involving flows between a city and a countryside without the help of a system of economic accounts, as do most of the other authors he cites, and does not appear to accept the export of services as a legitimate concept, jibing as he does that, "Most peasants and agricultural labourers probably valued good relations with the gods, but surely one should not regard the services of priests and soothsayers as economically relevant" (340). Of course they are, just as Delphi exported religious services, contemporary New York and London export financial services, Chicago exports medical services, and modern Greece exports tourism. Erdkamp does not cite Hopkins' 1980 article in his analysis of the consumer city, but he does implicitly formulate the payment of rents and taxes by rural residents to agents in a city as capital flows, and while Bang does not formulate the problem Hopkins addressed as the transfer problem of international economics, he does recognize the similarity between Hopkins' subject and Erdkamp's formulation of the flows between countryside and city in his defense of the consumer city typology. Erdkamp does go some way to providing a behavioral mechanism to underpin the consumer city concept but cannot resolve the basic problem with the urban taxonomic system of consumer city, manufacturing city, and other categories – that the concepts they codify as discrete types (even though apologists recognize what they call inexactitudes) are really measures on a continuum, particularly the size and identification of a city's export base, addressed in section 12.2.3.
- 3 The examples are overwhelming, but to offer just one instance of each case, consider Delougaz *et al.* (1967) for houses. For a remarkably informative excavation about an equally remarkable city, there are Peet and Woolley (1923), Frankfort and Pendlebury (1933) and Pendlebury (1951). On systems of cities, Adams (1981). And of course Pompeii and Herculaneum are preserved almost as caught by a camera (Zanker 1998; Jashemski 1979; Wallace-Hadrill 1994, 2011), while Akrotiri on Thera may eventually yield comparable evidence from 1500 years earlier (Doumas 1983). Ostia and Dura-Europos also are prominent among cities whose entire urban patterns are extraordinarily well preserved (Meiggs 1973; Hopkins 1979).
- 4 While the effects of crowding, sanitation and knowledge of diseases on ancient urban morbidity and mortality cannot be denied, there yet remains room for further analysis of the relationships between urban population growth and migration. Lo Cascio (2003) holds open the

possibility that migrants might not have been a major source of permanent population increase because of the predominance of males among them and the consequent low fertility rates of immigrant populations – a theme taken up by Erdkamp (2008, 238, 242–244); while Paine and Storey's (2003) demographic simulations suggest that migration could have stabilized the Roman age distribution fairly quickly after the Antonine Plague episodes of the second century C.E., although the Leslie matrix, which forms the basis of their analysis, routinely uses only female populations whereas the migration into Rome was always tilted toward males.

- 5 Postgate (1992, 66, Figure 3:13) reports house-sale tablets in Mesopotamia beginning in the Early Dynastic II period (2600 B.C.E.) through the Old Babylonian period (1700 B.C.E.), implying the existence of some kind of housing market; surely lower income city residents rented accommodations. Nevett (1999, 74) cites several studies reporting house sales in the fourth-century B.C.E. Greek city Olynthos, while Demosthenes 27, "Against Aphobus I," in Murray (1936, 1–53) is a lawsuit involving the inheritance of a house in fourth-century B.C.E. Athens. The Roman housing market has been well-studied. Wallace-Hadrill (1994) focuses on the owner-occupied market at Pompeii, and Pirson (1997; 1999, esp. 175) addresses the rental market in that city. Frier (1977) describes the rental accommodations at Ostia in the early imperial period and characterizes the rental market as economically inefficient, with its middlemen, delayed rent payments in leases, but does not analyze reasons for the specific forms of arrangements that existed. Frier (1980) details legal aspects of the Roman rental markets.
- 6 Meiggs (1973, 274–277) notes that most of the industry in Ostia, from the second-century B.C.E. through the fourth-century C.E., seems of small scale. Nonetheless, there were some exceptionally large buildings that apparently were warehouses for the distribution industry, and two especially large bakeries, one measuring 9,950 square meters in ground floor space, the other even larger. One deposit of jars in a warehouse had a capacity somewhat over twenty thousand gallons, most likely for wine or oil. He suspects the bakeries may have supplied Rome; that is, that bakeries composed an export industry in Ostia. (However, domestic ovens passed out of general use in Italy before the end of the Roman Republic, leaving bread a purchased consumption good for households.)
- 7 More specifically, according to location quotients, which are a measure of local concentration in an industry relative to the national share of employment in that industry. The measure is $L = (e_i/e)/(E_i/E)$, where e_i is employment in industry i in some urban area, e is that city's total employment, E_i is national employment in industry i , and E is total national employment. A value of L much greater than 1.0 generally is considered to imply that the city probably is exporting the output of that industry, whereas a location quotient substantially less than 1.0 is interpreted as indicating importation of the good (or service).
- 8 Owens (1991, 20) notes that, "Further away from the city center [of Classical Athens] ... houses become more spacious ..." Continuing more generally, he finds that "Certain areas of a city were more attractive than others. In these areas competition for space was often fierce ..." (26). He notes, on Delos, the clustering of commercial activity around what he calls "the city centre" near the port and the greater crowding of the residential district to the south of the harbor, toward and around the theater, than in that to the north of the sanctuary area, around Skardhana Bay (22–24).
- 9 Meiggs (1973, 273) notes that the excavations of Ostia have revealed a number of shops that did not have living quarters above them, indirect evidence that at least some of the daily workforce was commuting.
- 10 That is, open to migration flows, which fix utility at the level that can be attained elsewhere.
- 11 Not open to migration, hence having utility determined endogenously.
- 12 Of course, no one believes in the prevalence of one-person households, now or in antiquity. We could say rather arbitrarily that there are n people per household, which is simply a uniform household size greater than one, and the population density gradient would be shifted vertically. We could assume alternatively that households vary in size within the city but the size variation is spatially random. However, thinking about household size as a choice variable related to resources and opportunity costs as we did in Chapter 10, that approach to distributing households of different size across our model city can lead us in either of two directions: (i) since utility is the same everywhere in the city, resources and opportunity costs must be the same everywhere and the assumption of uniform family size everywhere may not be too bad; or (ii) consider several categories of household, distinguished by income, wealth, and family composition where we've been thinking until now of identical households. The latter line of thinking leads us to the construction of bid-rent curves for

- each class of household, with different types of household outbidding others for residential sites at different locations relative to the city center. Although we haven't used the bid-rent approach in our treatment of cities, we have introduced it in Chapter 11 to study the location relative to a fixed attraction site of different economic activities. Hochman and Ofek (1977) studied this problem by introducing different family sizes and compositions through multiperson budget constraints inclusive of the value of time.
- 13 Temin (2006, 44–45) observed that the very rare (and possibly unreliable) wages reported for Rome during the first and second centuries C.E. were higher than wages for comparable occupations in Egypt at the same time by a factor of three or four, noting the tendency in contemporary developing countries for wages in the center and large cities to be higher than those in the provinces and the countryside. The same can be said for contemporary industrialized countries as well.
 - 14 The basic relationship of the rank size rule relates the number of people who can be served by a single person in the smallest town: $p_1 = k(r + p_1)$, where r is the number of rural households, p_1 is the number of service workers in the town, and k is the number of service workers required to serve one household. Rearranging yields the expression for the population of the smallest town as $p_1 = rk/(1 - k)$, in which k is a small number, around the magnitude of 0.01 to 0.02. The term $k/(1 - k)$ can be called the urban multiplier. Any city of rank order n , with population p_n , serves s satellite cities, the total population of which, plus the surrounding rural areas is P_n . Then $p_n = kP_n$. But an n th-order city also serves as the center of an $n - 1$ st order market territory, so its total service population is $P_n - sP_{n-1} + P_{n-1} - p_{n-1} + p_n$. Making substitutions from the previous relationships, this last relationship, after some rearrangements, yields $P_n = [(1 + s - k)/(1 - k)] \cdot P_{n-1} = [(1 + s - k)/(1 - k)]^{n-1} \cdot P_1$. Now, going back to the population served by the smallest town, $P_1 = r + p_1 = r/(1 - k)$. Substituting this into the previous expression, the total population served by the n th order place is $P_n = [(1 + s - k)/(1 - k)]^n \cdot (r/(1 + s - k))$, and the population of the n th order place alone is $p_n = [(1 + s - k)/(1 - k)]^n \cdot [kr/(1 + s - k)]$. The last expression shows that city size is an exponential function of rank in the hierarchy.
 - 15 The exposition here follows the 1988 presentation, particularly Chapter 2, most closely. An additional, interesting analysis of city systems from the perspective of international trade theory is Hochman (1977), which has remained a fugitive manuscript but is widely cited.
 - 16 Thus the degree of industry-level increasing returns to scale first increases, then begins to peter out, leaving the inframarginal returns to scale intact but failing to confer further improvements over constant returns to scale once a certain city size has been reached.
 - 17 Henderson himself calls the entire term $N^{-\delta}$ a "spatial complexity" factor, a means of introducing the consequences of space without directly modeling it, which would enormously complicate the model.
 - 18 You can obtain the cost function by (i) taking first-order conditions for profit maximization for each production function, which yields the demand function for each factor; then (ii) substituting the expression for quantity demanded of each factor (which will be in terms of a price and technical production parameters) back into the production function; the quantity produced cancels out of the left-hand side of the production function, leaving the price of the good there in its place.
 - 19 Including the distribution of site rents, which are retained within the city. Income from capital is assumed to go to capital owners who reside outside the city. These are modeling conveniences, not expressions of beliefs in the locational structure of income receipts. Distributing income from capital across different types of cities, which have different costs of living, complicates the allocation of capital.
 - 20 The symbol Π is the multiplication version of the symbol Σ ; it tells us to multiply together the realizations of the subscripted variable. In this case, we multiply together the (appropriately subscripted) traded good prices.
 - 21 For the reader who wants completeness, $c_3 \equiv c_0^{-1/\alpha} \alpha^{-1} [f - b\beta(1 - \gamma)](f + \gamma b\beta)^{-1}$.
 - 22 For those readers who require the information, $E_1 \equiv E[f/(f + \gamma b\beta)]^{f+\gamma b\beta} [\beta^\beta B(1 - \beta)^{1-\beta} D^\beta (\gamma b\beta/f)^{\gamma\beta}]^b$ and $c_0 \equiv A^{-1} \alpha^{-\alpha} (1 - \alpha)^{\alpha-1}$.
 - 23 Henderson (1988, 225) found a 1% increase in city size accompanied by about a 0.50% increase in the nominal wage in the United States and 0.63% in Brazil in the 1980s.
 - 24 Once again, for completeness of detail, $c_4 \equiv (1 - \alpha)(f + \gamma b\beta)/[f - \beta(1 - \gamma) - \alpha(f - b)]$. Notice that k represents the citywide capital-labor ratio, K/N , while k_0 represents the capital-labor ratio in export production – the only production process in the model that uses both capital and labor, if we separate the intermediate production of sites from the final production of housing.

- 25 Some readers may be thinking at this point that it is absolutely absurd to think that we could ever recover the value of a purely conceptual parameter such as Cobb–Douglas production-function coefficients from ancient evidence. In one sense this certainly is correct: we could never obtain data to estimate these coefficients numerically. In a more important sense, however, it probably is within the realm of possibility to think that we could examine the evidence on ancient production technologies of different products and decide which processes seem to have required more labor per unit of output, which itself is a powerful bit of information.
- 26 Remember from Chapter 8 that steady-state growth means that labor *and* capital grow at the same rate, keeping the capital-labor ratio constant.
- 27 Classical Athenian income supplements to low-income residents of Athens in the Classical period would have had the same effect on the population of Athens relative to other Greek cities and the same effect on the existence of smaller Greek cities. During the fifth and fourth centuries B.C.E. various social benefits were available to Athenian citizens: means-tested disability benefits for the impecunious; maintenance of children of soldiers killed in battle; non-means-tested festival money, and occasional grain distributions during periods of hardship (Carey 2000/2001, 39; Migeotte 1997, 38–39). Loomis (1998, Chapter 13) cites reports of welfare and support allowances and distributions, noting some grain distributions and disability relief for the impecunious of 2 obols per day by 322 B.C.E. Lysias 24, “On the Suspension of the Benefit of the Disabled Man,” while contested by some scholars, depicts a scenario of a poor, disabled, fourth-century Athenian that is difficult to dismiss (Wolpert and Kapparis 2011, 65–72). According to Finley (1980, 90), “... Athens in the fifth and fourth centuries B.C. went much further than any other state in providing ways of supplementing income.” Altogether these benefits probably did not rival the public benefits available to some Roman citizens at Rome, they still could have had the effect of adding to the attractions of Athens within Attica.
- 28 I say “valued” amenities because if, say, a pleasant climate is enjoyed but consumers are so poor that they are unwilling to give up other items of consumption to have more of it, climate will command no rental price in local land values and wages.
- 29 I recognize that the phrase “entire country” has a modernist ring, but the idea that some towns and cities have routine interactions with a smaller geographical area than do some of the cities that form part of the small cities’ urban interaction partners surely has applicability in much earlier times.
- 30 Even if you don’t like or morally approve of a government, its expenditures are public expenditures and the things they provide are public services, even if some or many of those services are provided much less efficiently than other methods of provision could deliver. State-delivered religious propitiation of the gods is a public service, provided by state-supported priests or a priest-king. The fact that the ancients wouldn’t have seen it that way should not affect our own surgical techniques today, any more than we have ceased using ancient and medieval chemical and physical theories and adopted more powerful theories that account for the same phenomena, and other related ones, more reliably.
- 31 Some readers may reflect that surely these governments made no interest payments to any lenders for these construction projects, and we would agree wholeheartedly that those payments need not have taken such a form. However, the foregone and deferred private consumption of goods and leisure required to construct these projects has exactly that form. Contemporary city residents ultimately pay the interest on their cities’ debts through their taxes; their predecessors in the cities of antiquity would have paid these real resource costs somewhat more directly, through *corvée* labor or shares of their crops.
- 32 Another type of (impure) public good is called the “club good.” If one were inclined to be taxonomic, one could designate local public goods a subset of club goods (Cornes and Sandler 1986, Part IV, esp. Chapters 10–13). A club good is a congestible public good; that is, there is a minimum feasible size to the good, and at that size multiple people can consume it without detracting from one another’s consumption. Eventually congestion sets in – the facility becomes crowded, users have to wait in line to consume, and so forth. A club could be formed by consumers to provide such a facility for themselves, the optimal size of the club being determined by the condition that the price of membership should be the cost to members of the value of the congestion losses imposed by the last member. Local public goods have this congestibility characteristic but also have a spatial dimension, providing benefits only to residents of a particular area.
- 33 Of course, we could find the occasion when the royal army got itself holed up within the walls of one particular city during empire-wide hostilities. This isn’t what we’re talking about.

- 34 For overviews of analyses of the Tiebout model, see Rubinfeld (1987, section 2, 581–601) and the studies in Zodrow (1983). McDonald (1997, 262–272) provides an accessible comparison of the Tiebout and median-voter models.
- 35 If everybody understates his or her preference, there is just a bigger deficit in suboptimal provision.
- 36 They could differ only in income.
- 37 The “outside grant” in the Graeco-Roman world includes euergetism and, in Athens, the liturgies mentioned in Chapter 6, section 3.5.
- 38 However, if the valuation of local public goods purchased with the property tax revenue is also capitalized into the values of houses as amenities, this result is seriously compromised, and there could be close to a wash.
- 39 A reasonable elasticity, or the $\% \Delta V / \% \Delta t$, would be in the range of -0.25 , using a price elasticity of demand for housing services of -0.7 and a short-run supply elasticity of housing of 0.2 .
- 40 The elasticity of the land price with respect to a tax change is a factor of the inverse of land’s share in the housing cost or production function times the tax elasticity of the housing price. A reasonable range of values for land’s share in housing costs is 0.1 to 0.2 .

Natural Resources

Natural resources generally occur without the intervention of man, in contrast with such goods as housing, clothing, and food. Some of them exist in a fixed quantity within any time period of relevance to humans. Several hundred millions of years could see the development of some more metal ores or limestone or granite or coal, but for human purposes this regenerative period is irrelevant. These resources are called exhaustible or nonrenewable resources. Other resources, mostly biotic, but not exclusively, have rejuvenation periods that are short enough to replenish their stocks during human planning horizons. Fish, birds, game animals, timber, grassland, soil fertility, and groundwater are such renewable resources.

The distinction between these two categories of resources can blur because discoveries of new stocks of many exhaustible resources can increase their known reserves by more than current mining or other extraction reduces them. Their existing reserves can actually increase over time. Technological changes in how they are used can also make their effective reserves increase if their requirements per unit of final good fall or the resources required to extract them fall. Renewable resources, on the other hand, can be effectively “mined” – harvested at a rate in excess of their growth rates until individual stocks are

exhausted or entire species are made extinct. All resources *can* be exhausted.

The efficiency of the allocation of exhaustible resources over time depends on institutional structure. Complete efficiency of intertemporal resource allocation requires a complete set of forward markets, which is never found today, and was not found in antiquity, and consequently resource markets tend to be unstable and to deplete or harvest, as the case may be, at nonoptimal rates. The models developed to analyze optimal depletion and harvesting begin with the assumption that the resource managers know the initial size of the resource stock, which generally is not the case, even with sophisticated, contemporary prospecting and population estimation technologies. Despite the empirical falsity of this assumption, the importance of this knowledge, as well as the path of the future demand price for the resource, will focus resource managers’ best attentions on learning the former and forecasting the latter. Nonetheless, estimates of resource availability may increase or decrease unpredictably, and prices may fluctuate correspondingly as well, from technological changes and the appearance of substitutes.

Section 13.1 develops the theory of optimal depletion for exhaustible resources, since exhaustion of fixed resources, even allowing for

discoveries, is in the cards for these resources. The best strategy people can adopt is to mine them at the rate that allows their consumption to be the highest for the longest time. Section 13.2 turns to the theory of optimal use of a renewable resource: since the stocks of these resources can be regenerated, it is possible to find the harvest rate that will be sustainable – essentially forever, or at least for discountable planning horizons – and give people the highest consumption level. While most exhaustible resources can be effectively claimed as private property, many renewable resources elude both definition and enforcement of property rights and consequently experience over-use attributable to their open-access character. The final section looks at how we might infer resource scarcity from observable indicators.

13.1 Exhaustible Resources

The basic economic theory of exhaustible resources is the theory of depletion, which customarily is presented in a pretty unrealistic setting because to do otherwise would obscure the basic results. Accordingly, we present that theory in subsection 13.1.1, then proceed to make its setting more realistic through the following four subsections, adding different deposits, uncertainty about all kinds of things, prospecting (more formally dubbed exploration), and the possibility of monopolization, which with fixed-location, geographically random resources is not a modern possibility restricted to petroleum.

13.1.1 The theory of optimal depletion

The quantity of an exhaustible resource is fixed, so present consumption involves an opportunity cost in the form of the value that might have been derived at some future date. This causes the first difference between efficiency conditions in exhaustible resource extraction and ordinary production: market price must be higher than marginal cost, as shown in Figure 13.1. Second, intertemporal efficiency suggests intuitively that the tradeoff between present consumption and future consumption ought to be the same across all the time periods for which people can plan.

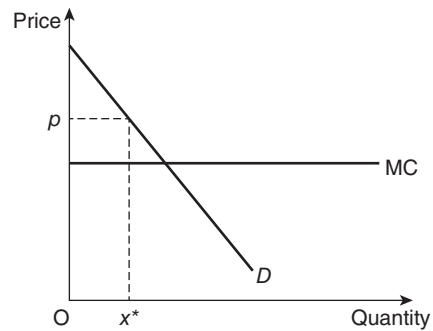


Figure 13.1 Market price > marginal cost in optimal resource depletion.

We will see exactly how both of these conditions emerge from a simple maximization problem encountered in planning operations in, say, a mine.

In this maximization problem we could think of the owner of a single mine planning his operations or of the collectivity of all mines exploiting known deposits of a particular mineral – the structure of the problem is the same. Once mined, this mineral can be sold for a price p_t in the current period (denoting time periods by t). Assume that the future price path is known; of course, it can't be, and we'll explore later the effects of uncertainty about it and of unanticipated changes in it. Next assume that the initial size of the mineral reserves (S_0 for the stock in time period zero) is known, as well as the amount that the mining operation will leave at the end of the operational period (S_T for the stock in the terminal period). The director of mines wants to maximize the present discounted value of benefits from exploiting the mine, which means maximizing the difference between sales revenues and extraction costs in each period (profits), but since the profits arrive over time, he must account for the fact that earlier profits are worth more than later profits – for which he accounts by discounting future profits. The decisions he can make to accomplish this goal are the quantities of the ore he mines in each period, x_t – an entire time path of extraction rates. The following exposition is adapted from Dasgupta and Heal (1979, 169–175), Fisher (1981, 25–35) and Conrad and Clark (1987, 117–126).¹

Give the mine a simple cost function: costs are higher in any period the more of the mineral extracted, and a larger stock is easier to dig ore out of than a smaller one. That is, the cost function is $c(x_t, S_t)$, where the positive effect of the extraction rate gives $\Delta c/\Delta x_t > 0$ and $\Delta c/\Delta S_t < 0$ (or more compactly $c_x > 0$ and $c_s < 0$). So, each period, the profit is represented by $p_t x_t - c(x_t, S_t)$, and the profit of each future period by $[p_t x_t - c(x_t, S_t)]/(1+r)^t$, in which the value of t is replaced by the number of the time period (1 for period 1, 2 for period 2, and so on) – an expression we’ve seen plenty of times before. To get the total of the profits of the mine in all time periods we just add them up from period 0 to the terminal period, labeled T: $\sum_{t=0}^{T-1} \{[p_{tx_t} - c(x_t, S_t)]/(1+r)^t\}$, which the director maximizes by the appropriate choice of time plan of extraction rates, $x_0, x_1, \dots, x_{T-1}$.²

The mine owner (or director of mines) is constrained to extract no more over the mining period than the difference between the beginning and ending stocks. The extraction in period zero is the difference between the size of the initial stock before mining begins and the size of the stock at the beginning of the next period (which is equivalent to the end of the initial period): $x_0 = S_0 - S_1$. Similarly with period 1: $x_1 = S_1 - S_2$; and so on until we meet the penultimate period, in which $x_{T-1} = S_{T-1} - S_T$. Now, add all these up – both sides of these expressions – to get the overall extraction constraint that the director of mines faces: $\sum_{t=0}^{T-1} x_t = S_0 - S_T$. To formulate the maximization problem completely, we need to specify that the initial and terminal stocks are fixed in magnitude: $S_0 = \bar{S}_0$ and $S_T = \bar{S}_T$, where the bars over the letters indicate that those quantities are exogenous, and not to be determined by the maximization process.

To find the relationships between variables that the maximization will entail, we set this problem up more formally than an ancient director of mines would have done – we form the Lagrangean expression that we have used so often before: $\mathcal{L} = \sum_{t=0}^{T-1} \{[p_{tx_t} - c(x_t, S_t)]/(1+r)^t\} + \sum_{t=0}^{T-1} \lambda_t (S_t - S_{t+1} - x_t)$. (I ignore the constraints on the initial and terminal stocks, because they don’t show anything particularly interesting.)

Remember that λ_t is the Lagrange multiplier, and that its interpretation is the value of a slightly larger value of the constraint – in this case, the value of another unit of the unextracted ore in the ground, a slightly larger initial stock of the mineral. This is the opportunity cost of taking a unit of ore out of the ground and using it this period rather than leaving it for later. Sometimes this value is called the “royalty,” sometimes “rent,” sometimes the shadow price of unmined stock. It turns out to be a very important element in the economics of exhaustible resources.

Next, remember the first-order conditions for a maximization – we change the value of each factor that varies in the maximization over some range until we find the maximum value of the Lagrangean function; the value of the variable that gives that value is the optimum value of that variable. We’ll look at the first-order condition for three variables. The mine director’s variables to adjust are the extraction rates in each of the time periods, x_t . Adjusting the size of the stock at the beginning of each period is equivalent to changing the previous period’s extraction rate, but gives us a different perspective on the optimization. Adjusting the value of the Lagrange multiplier, λ_t , finds the profit-maximizing royalty, or scarcity value of ore in the ground.

The first-order condition for the extraction rate in each period shows the relationships caused by finding the profit-maximizing extraction rate in each period: $p_t = \lambda_t(1+r)^t + c_x$, which is the price > marginal extraction cost relationship noted in the introduction to this section. Along the optimal (that is, efficient) depletion path, the sale price of the extracted mineral should equal the sum of the marginal extraction cost and the royalty, or opportunity cost. You can see that if mining technology changes rapidly enough, the market price, p_t , could fall for a while even though royalty continues to increase.

The first-order condition for stocks tells us that the royalty rate must rise over time to account for increasing scarcity value of ore in the ground as more ore is extracted: $\dot{\lambda}_t/\lambda_t = r + c_s/\lambda_t$, in which the “dot” notation means “change in” ($\dot{\lambda} = \lambda_t - \lambda_{t-1}$), and dividing by λ_t converts the difference into a percentage difference, or a

growth rate. Thus the royalty would grow at the rate of interest (the discount rate, r) if extraction cost were not affected by stock size. As the mining cost is specified, the royalty grows at a rate somewhat smaller than the interest rate. When the stock is large – during early periods of mining – the royalty is a small proportion of p_t , so while the royalty rises at roughly the rate of interest, p_t won't; most of p_t is extraction cost in early periods. When the stock is large, its exhaustibility doesn't excite much interest. Over time, as the stock is depleted, the royalty component of price will come to dominate. Figure 13.2 shows the time paths of the market price and the royalty, under the assumption of a constant marginal extraction cost. At early times in the extraction program, the ratio p_t/λ_t is large, their values being separated by the marginal extraction cost, and the absolute difference between them stays the same over time, reducing the p_t/λ_t ratio.

The third first-order condition optimizes the rate of extraction – the amount of ore mined – in each of the periods and comes up with the simple relationship that the decrease in the stock equals the extraction in that period: $\Delta S_t/\Delta t = -x_t$. What will this time path look like? The remaining resource stock is getting smaller in each period, which will increase the unit cost of whatever is mined. Even if the marginal extraction cost doesn't stay the same in each period as we supposed for diagrammatic convenience in

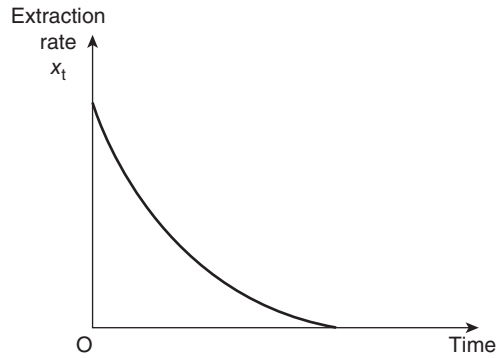


Figure 13.3 Time path of extraction rate.

Figure 13.2, the quantity mined will have to fall to remain equal to the difference between the market price and the royalty over time. Consequently the extraction path will look something like the decreasing curve in Figure 13.3, which should be compared with the increasing price and royalty of Figure 13.2. By the terminal period, the increasing unit costs accompanying the declining stock drive the quantity of ore that can be mined profitably to an inconsequential level.

The availability of a substitute for a natural resource, at a price higher than the resource's extracted price in early years places a limit on its price increase over time, as depicted in Figure 13.4. The price of the resource will not immediately jump to the price of the substitute – sometimes called a backstop technology – but

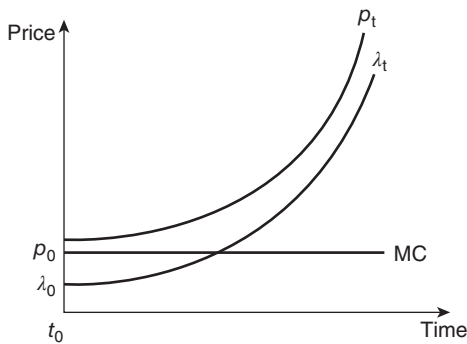


Figure 13.2 Time paths of market price and royalties.

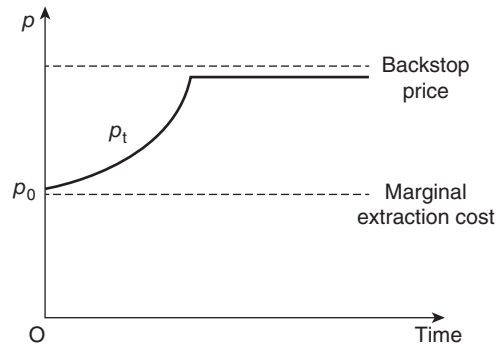


Figure 13.4 Effect of a substitute on price of a resource.

will approach it at its ordinary rate of increase, roughly the rate of interest. In many cases, the backstop resource will not become profitable to use, at least in most technical opportunities, until the price of the exhaustible resource has progressed along its time path and reached the level of the backstop.

13.1.2 Different deposits

Under competitive conditions, deposits with different qualities of ores, or other sources of differences in extraction costs, will not be mined at the same time. How can we reconcile this contention with the empirical observation that many mines operate contemporaneously? The key is in the qualifier “competitive.” We have not addressed the aggregate demand for this mineral but clearly the price is given to both of these deposits, which implies that together they supply a small enough share of the output of this mineral that their outputs don’t affect the market price, which rises as a function of depletion over all mines. We could interpret deposits 1 and 2 below as typical mines of two qualities that exist across all mines, or deposit 1 as the lowest quality of an array of mines that can produce at the given market price and deposit 2 as the highest quality mine of an array of lower quality mines that cannot produce given the prices during the early portion of our time period.

Within these market circumstances, the claim is easy to show. Suppose we have two deposits, deposit 1 having richer quality ore and hence lower mining costs, deposit 2 being of lower quality. For mine (deposit) 1, the price-royalty relationship is $p_{1t} = \lambda_{1t} + m_1$, where m is the marginal extraction cost, which we can assume to be constant over time without losing anything important. For mine 2, that relationship is $p_{2t} = \lambda_{2t} + m_2$. Lower costs in mine 1 mean that $m_2 > m_1$. Now, obviously, the two mines produce the same thing, so $p_{1t} = p_{2t} = p_t$.

As long as both mines continue to operate, the growth rate of the scarcity value (royalty value) of the ore is the same in both: $\dot{\lambda}_{1t}/\lambda_{1t} = \dot{\lambda}_{2t}/\lambda_{2t} = r$. As long as the two mines face the same market price and their extraction costs differ, their unit royalty values can’t be the same, even though

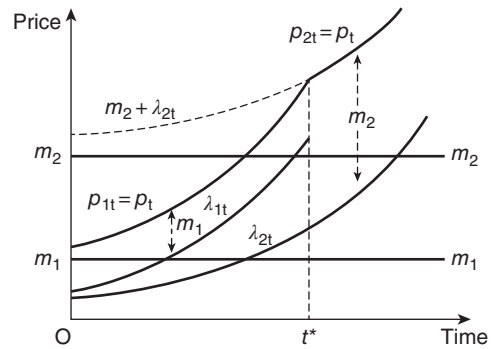


Figure 13.5 Resource extraction from different deposits.

they increase at the same rate. During some initial group of time periods, during which p_t , λ_{1t} , and λ_{2t} are all rising, $\lambda_{1t} + m_1 < \lambda_{2t} + m_2$, and mine 2 can’t deliver the mined ore at a profit. However, having started off at a higher royalty level, eventually the royalty plus extraction cost at mine 1 rises to equality with the royalty plus extraction cost at mine 2, beyond that time exceeds it, and, having become the high-cost mine facing a fixed market price, shuts down in favor of mine 2. Figure 13.5 depicts the progress of price and royalties. While mine 1 is exhausted in an economic sense at the switching time, t^* , it may have plenty of ores that would become exploitable again with technological change or an unanticipated increase in demand.

This assessment should not be taken as logical proof that Saudi Arabia and Iran can’t pump oil at the same time. It is an examination of the consequences of a competitive market structure in which some deposits can supply ore at or below the prevailing price and others cannot. To think about a market structure in which several deposits of different quality are exploited simultaneously, we must replace the fixed price (actually a price path) with an inverse demand curve (price as a function of quantity on the market, rather than quantity demanded as a function of price), find the quantity that can be sold in each time period, and allocate that production among the deposits according to extraction cost. The marginal mine will be the one with the lowest quality reserves (highest extraction cost per unit of ore), and its

royalty during the first period it can operate will be zero.

13.1.3 Uncertainty

We've assumed perfect knowledge about several key parameters of mining that generally are only imperfectly known, even with contemporary technologies and market institutions. We can categorize these as demand uncertainties and supply uncertainties. Uncertainties about demand can take several forms. First, prices farther in the future surely will be less well known – mine owners will be less confident of their predictions about them – and risk-averse producers will shift extraction toward the present, just as if the discount rate had risen. Alternatively, suppose that the price in each period is a random variable (known only within a range by the producer) related positively to the volume of ore mined in the same period – the more you mine the greater the uncertainty (variability) in price. As price rises and the quantity mined falls over time, the variation in price, being proportional to output, will also fall. Risk-averse producers will shift depletion to the future, when the variability of prices is smaller – just as if the discount rate had fallen.

Uncertainty about the appearance of a lower cost substitute at some future date generally would lead a mine owner to accelerate his depletion. The appearance of the substitute will cause the price of the currently mined resource to fall. The risk of expropriation of a deposit will accelerate depletion as well, because on expropriation, the value of the deposit to the (former) owner falls to zero. It is likely that toward the end of the Late Bronze Age or the beginning of the Early Iron Age, when some of the technical possibilities of iron were becoming better known, uncertainty about the ability of the new metal to supplant bronze in most of its uses would have accelerated copper mining, even though mine owners were unfamiliar with the theory of optimal depletion. They could look down the pike and see competition and understand its potential, even if, as in the case of iron, that potential was not fully realized. Despite being known and occasionally smelted as early as the Early Bronze Age, iron production

apparently remained too costly relative to bronze production for many years – we could even say centuries or millennia – for iron to replace bronze except in the more highly valued uses, such as weaponry (Waldbaum 1980).

The principal type of supply uncertainty is lack of knowledge about the size of the stock. A risk-averse producer would want to avoid running out of exploitable reserves unexpectedly because his extraction plan would not have maximized the value he could have obtained from the total stock. That is, a producer who ran out of reserves unexpectedly would want to go back in time and revise his extraction plan downward – but of course it's too late by then. As a more intricate set of uncertainty relationships, a high discount rate will retard investment in both extraction capital and exploration and development of new deposits. This situation will lead to an initial, high extraction rate directly because of the high discount rate, followed by a period of slower mining caused by reduced investment in exploration in early years.

13.1.4 Exploration

Exploration uses real resources to expand the stock of a depletable resource and thus lower the extraction cost. If the prospecting finds new resources, owners of currently operating mines will expect prices to fall in the future and will expand their own current depletion, hastening the price decline to the present. An interesting relationship emerges from studying the mine owner's maximization problem when she engages in exploration. Use the same extraction cost function, $c^1(x_t, S_t)$ and add an exploration cost that is a function of the new reserves found with the effort, d_t , $c^2(d_t)$, in which cost increases as more reserves are found: the marginal cost of exploration being $c_d^2 > 0$.³ Then the profit maximization problem becomes $\max_{\{x_t, d_t\}} \sum_0^T [p_t x_t - c^1(x_t, S_t) - c^2(d_t)] / (1+r)^t$, and the constraint relating the change in resource stocks to extraction becomes $S_t - S_{t+1} = d_t - x_t$, and we see that the resource stock could actually increase over some time period.

From the first-order condition governing the optimal rate of depletion, the relationship

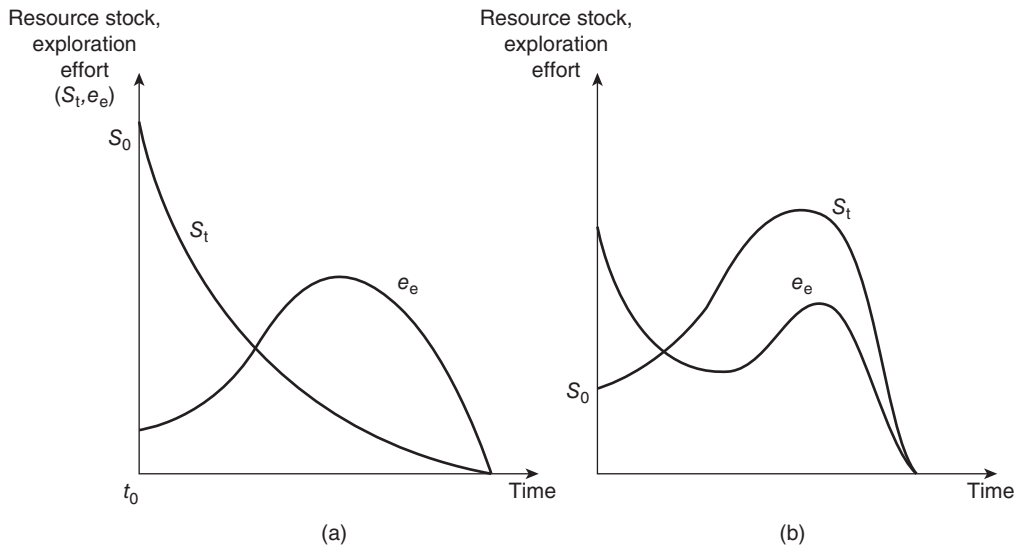


Figure 13.6 (a) Time paths of resource stock and exploration, initially large stock. Adapted from Conrad and Clark 1987, 135, Figure 3.3a. © 1987 Cambridge University Press. (b) Time paths of resource stock and exploration, initially small stock. Adapted from Conrad and Clark 1987, 135, Figure 3.3b. © 1987 Cambridge University Press.

between price, royalty and marginal extraction cost remains the same: $p_t = \lambda_t + c_x^1$. The first-order condition telling the mine owner when she is doing the optimal amount of prospecting is $\lambda_t = c_d^2$: in each time period the royalty value is equal to the exploration cost.⁴ The royalty value is the shadow value (or shadow price) of a unit of stock remaining in the ground, unmined, which in equilibrium equals the cost of finding another unit to replace it. Without prospecting, the interpretation of royalty was the value of keeping an additional unit of the ore in the stock; when it's possible to add another unit to the stock, the benefit of doing so must be balanced by the cost of doing so. Figure 13.6 shows several possible patterns of exploration effort and depletion. Panel (a) shows a case of an initially sizeable resource stock, S_0 , which induces low exploration efforts, designated by e_e . The resource stock is depleted continually to exhaustion, with exploration picking up during the period of serious depletion. In Panel (b), initially small, known reserves prompt early exploration efforts which discover new deposits. The known resource stock increases during the early period of mining, and exploration efforts decline in response to the attendant price decline. However, the resource eventually

is depleted, and exploration efforts pick up again in later stages of depletion as the resource price rises, but are unable to locate further reserves to avoid exhaustion (Conrad and Clark 1987, 135).

As noted earlier, discoveries will reduce the present market price of a resource through the current reactions to expectations of future price decreases caused by the discoveries. With a series of discoveries, an exhaustible resource can have a temporal price pattern that looks like a series of waves, as in Figure 13.7. Price rises at the rate

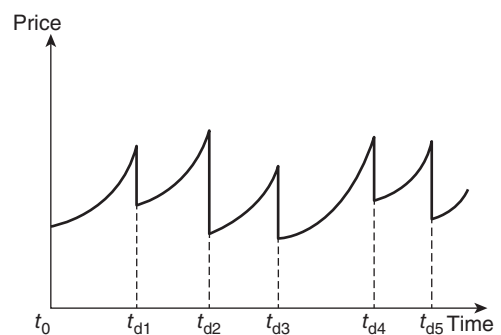


Figure 13.7 Resource price over time with sequential discoveries.

of interest, then upon discovery of new reserves, drops precipitously, only to begin rising again at the rate of interest. Further new discoveries drop the progress of the rising price and let it start all over again. Over some lengthy period of time, say several centuries or more, you can see that the price need not increase much at all, if any, over its early levels. Incomplete observations over time that yield the impression of no clear trend in the price of an exhaustible resource could be explained at least in part by such a pattern of periodic discoveries.

13.1.5 Monopoly

Monopolists' profit maximization problem is the same as the competitive mine's, but their first-order condition for optimal extraction differs because their mine's output influences the market price of the ore and they must take this into account in their extraction plan.⁵ Consequently their version of the price-royalty relationship is that marginal revenue minus marginal extraction cost equals royalty: $p_t + x_t \Delta p_t / \Delta x_t - c_x = \lambda_t$, in which the expression $\Delta p_t / \Delta x_t < 0$ represents the depressing effect of additional (marginal) ore on the price of ore and the sum of the first two terms is marginal revenue.

Royalty behaves over time as it does for a competitive mine, but whether extraction is faster or

slower than in the competitive mine depends on the behavior of the relationship between price and marginal revenue over time. If the elasticity of demand for the ore is larger (in absolute magnitude; remember that it is negative) for smaller quantities, the monopolist's price path over time will be flatter than that of a competitive mine. The monopoly price path will be above the competitive price path during early years of extraction and lower in later years, as in Figure 13.8(a). Correspondingly, the monopolistic extraction rate is lower than the competitive extraction rate during the early mining years and higher later. If the elasticity is larger for larger quantities of ore extracted per period (lower prices bring the ore into competition with substitutes as the ore penetrates the markets for other products), the monopoly price path over time is steeper than the competitive price path, the monopolistic price is lower than the competitive price in early years and higher later, and depletion rates are faster early and slower later than the competitive depletion rate, as in Figure 13.8(b). This case yields a growth of price at a rate higher than the interest rate, which may not be sustainable.

Dasgupta and Heal (1979, 331) find an intermediate case more plausible empirically than either of these pure cases. Specifically, the demand elasticity for the ore is likely to increase as output rises and also as price rises, giving a minimum (least

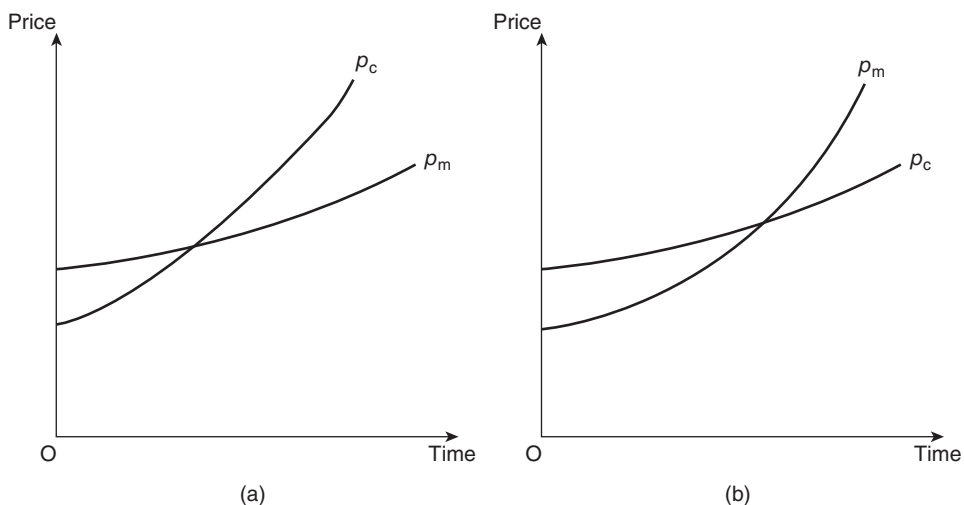


Figure 13.8 (a) Monopoly price path: more elastic demand for a smaller quantity of ore. (b) Monopoly price path: more elastic demand for a larger quantity of ore.

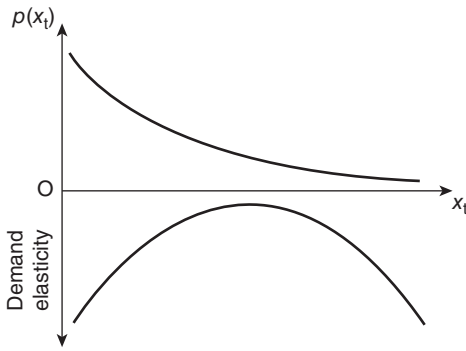


Figure 13.9 Monopoly price path: intermediate case.

elastic) value at some intermediate rate of both price and extraction rate, as shown in Figure 13.9. In any event, the essential point is that the monopolist restricts ore output to take advantage of less elastic demand, whether that elasticity increases or decreases as the annual extraction rate rises.

13.2 Renewable Resources

Renewable resources are regenerative but potentially exhaustible. While ownership of the typical exhaustible resource can be claimed and enforced, many renewable resources cannot be so regulated. Such unowned resources frequently are called common property or “the commons,” but “open access” is a preferable description because common property resources can be subjected to controlled use just as completely private resources, whereas overuse is to be expected in truly unowned resources. Nevertheless, optimal harvesting of an owned renewable resource can lead to extinction while exploitation of an open-access resource may reach a stationary, or steady, state short of extinction (Dasgupta 1982, 14).

The optimal use of a renewable resource has much in common with the optimal depletion of an exhaustible resource, but holding back on harvesting will allow the renewable resource to grow, without the necessity of using additional resources for exploration and development as in the case of the exhaustible resource.

Consequently, the first element in the economics of how best to use a renewable resource is the structure of the biological growth. Next we’ll take up the economics of the harvesting activities in a way parallel to our examination of the production and cost theory of mining, with the addition here that harvesting interacts with the biological growth independently of the ownership of the resource. Then we turn to the optimal harvesting problem in the case of an owned resource, and finally to the exploitation of an open-access resource, using a fishery as an example.

13.2.1 Biological growth

The biological growth law most commonly used is that growth – in terms of numbers of individuals, size, or both – is a function of stock size, following a logistic growth path as in Figure 13.10(a), which plots the size (or biomass) of a natural population over time.⁶ It reaches a maximum size asymptotically. The growth in any period is a function of the size of the stock (or population) at the beginning of that period, with growth increasing with larger stock size up to some point of maximum growth, then declining to zero at the largest sustainable population size, as in panels (b) and (c) of Figure 13.10. Since growth during any period is the amount by which the stock increases in the same time, $\Delta S_t / \Delta t \equiv \dot{S}_t = G(S_t, \Omega)$, where Ω represents any other influences on growth. When $\dot{S}_t = 0$, the population remains a constant size, and if that size is nonzero, that is, the population isn’t extinct, it is called stationary, and in both panels (a) and (b), $\bar{S}$ is the natural, long-run population size in the absence of human predation. The growth at population size $\bar{S}$ is the maximum sustainable yield: the maximum growth does not occur at the maximum population size.

The growth curve in Figure 13.10(b) allows a population to survive even at very small sizes, but members of the species represented in Figure 13.10(c) have difficulty finding one another for reproduction when the population falls below $\underline{S}$ and at that population, the growth rate becomes negative. These growth functions are quadratics: in the case of panel (b), $\dot{S}_t = aS_t - bS_t^2$, and in panel (c), $\dot{S}_t = -c + aS_t - bS_t^2$, in which

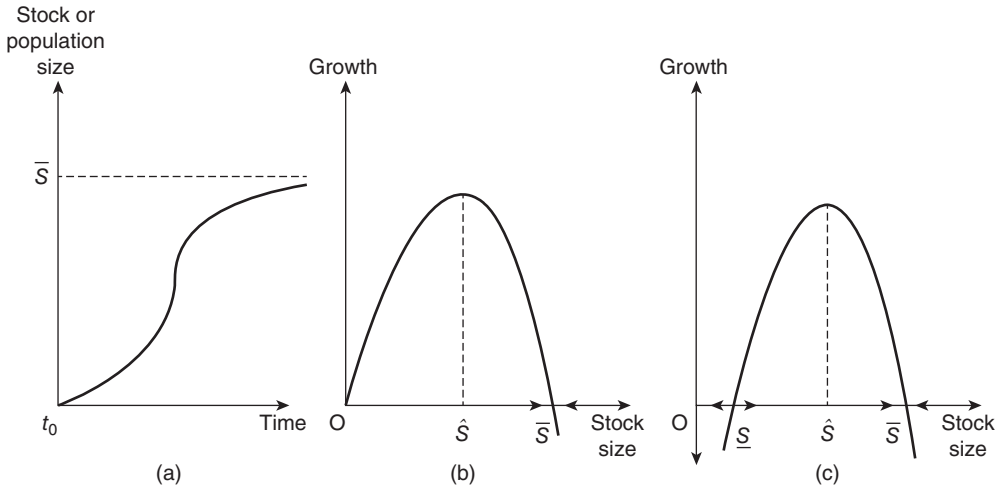


Figure 13.10 (a) Biological growth in stock of renewable resource. (b) Maximum sustainable yield: stock survives at small size. (c) Maximum sustainable yield: stock goes extinct at small size.

$a, b, k > 0$. For both types of growth, the population size that generates the maximum sustained yield is $\hat{S} = a/2b$. In panel (c), an initial population size between $\underline{S}$ and $\bar{S}$ will grow toward $\bar{S}$, as indicated by the directions of the arrows on the horizontal axis; if population gets larger than $\bar{S}$, its growth will become negative and population will fall back toward $\bar{S}$. If population falls below $\underline{S}$ it will die out. At both points, growth is zero, the population at $\bar{S}$ being stable, in the sense that a small departure from that population would set off forces returning the population to that level; that at $\underline{S}$ is unstable, in the corresponding sense that small departures would not return population to $\underline{S}$.

13.2.2 Harvesting

We introduce harvesting to this natural environment and the growth becomes $\dot{S}_t = G(S_t, \Omega) - x_t$, where x_t is the quantity harvested. If $x_t = G(S_t, \Omega)$, $\dot{S}_t = 0$, and we have a sustainable harvesting rate. The following exposition is adapted from Dasgupta and Heal (1979, 119–126), Fisher (1981, 79–86), Dasgupta (1982, 120–130), and Anderson (1977, 22–40). The interesting questions, of course, are whether such a rate can be reached, and under what conditions. We

introduce a harvesting production function for a privately owned renewable resource in this subsection, reserving the specification of open-access production to subsection 13.2.4.

The harvesting rate per period is an increasing function of both the stock of the renewable resource at the beginning of the period and the “effort” expended, an amalgam of labor and equipment used in harvesting by our operating unit: $x_t = f(S_t, e_t)$, $f_S > 0$, $f_e > 0$, $f_{SS} < 0$, $f_{ee} \leq 0$, where the last two notations indicate that the increments to catch permitted by a larger stock decrease with stock size (that is, we have a shape of a harvest, or yield, function similar to the growth functions of Figure 13.10), and the marginal returns to increasing effort also might have such a declining pattern in additional effort or they might be linear in effort – that is, constant cost. Additionally, we have the two basic conditions that $f(0, e_t) = f(S_t, 0) = 0$, which simply says that with no stock or with no effort, we get no harvest; the first implies that we get no further yield from an extinct population and the second that harvesting is not costless. Figure 13.11 plots yield (also called harvest and catch) against stock (population) and effort, in two panels. The yield axis is labeled “growth” also because both harvest and growth bring changes

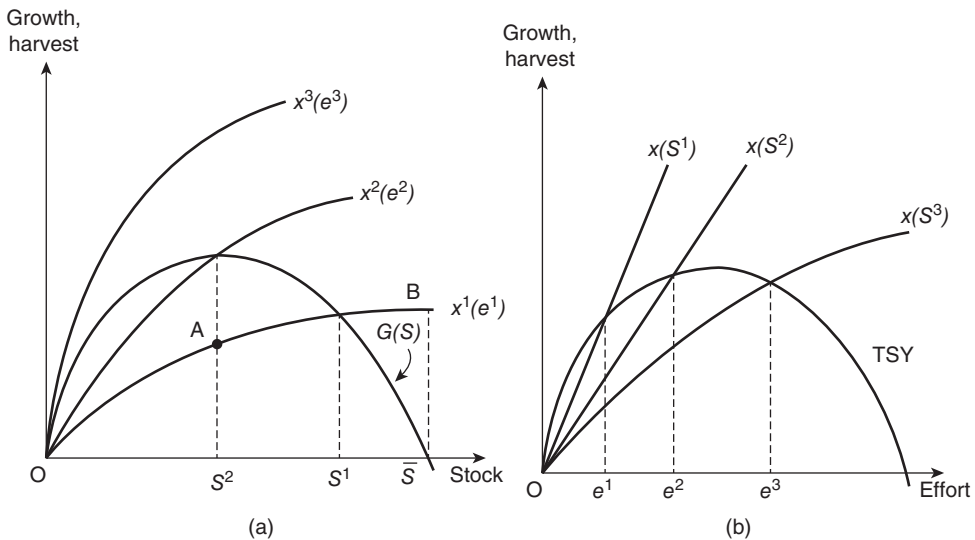


Figure 13.11 (a) Harvest-stock relationship. (b) Harvest-effort relationship.

in stock, which can be distinguished only in their causes. In panel (a), curve G is the given population growth curve, and $x(e^1)$, $x(e^2)$, and $x(e^3)$ are yield curves for three different levels of effort, $e^1 < e^2 < e^3$ (we use superscripts to distinguish different magnitudes of effort from the time period in which the effort is applied, noted with subscripts). Effort level e^1 is consistent with a stationary population size of S^1 . At point A, this effort on population S would yield a catch smaller than the natural growth of the population, leading to an increasing population. Applied to the maximum population size, S , the yield at B is far in excess of growth, and the population would decline. From either direction, the population will converge on S^1 , given the application of effort e^1 and the growth characteristics. If effort level e^2 is chosen, a larger harvest can be brought in each period, but from a smaller population, S^2 , at which growth is larger. Increasing effort to e^3 will extinct this population – or exhaust a stock. The drawdown of biomass with this quantity of effort yields a harvest larger than the natural replenishment through growth for any population or stock size, and this population is doomed to extinction.

In panel (b) the quadratic curve is the total sustainable yield as a function of effort, given different population sizes. You can think of it as the long-run production of the resource

husbanding enterprise. The curves $x(S^1)$, $x(S^2)$, and $x(S^3)$ are short-run production (yield) functions for particular levels of population $S^1 > S^2 > S^3$. An increase in effort from e^1 to e^2 reduces the equilibrium population but raises steady-state yield by moving from population S^1 to the smaller population S^2 , which has a larger growth. Further increase in effort to e^3 reduces the stock to S^3 and gets essentially no increment in yield for the additional effort. In this diagram, the changes in effort are the exogenous forces, not shifts in technology; the changes in effort, by changing population size, alter the effort cost of catch but nothing fundamental in the technology.

We can invert the production function to get a function for effort: $e_t = c(S_t, x_t)$, $c_S < 0$, $c_x > 0$: the effort required to bring in a harvest of x_t is smaller the larger is the size of the stock or population – the fish or deer are easier to find when there are more of them – and for a given stock size, extracting a larger harvest will require more effort. We can express yield cost as a function of resource stock, $C(S_t)$, $C_S < 0$, meaning that harvesting any particular yield from a smaller stock or population is more costly. Figure 13.12 plots two yield cost curves as functions of population size ($S^1 > S^2$). Line px plots the total revenue of any yield size, given a price p of the harvested resource. At a yield of x_t^1 the

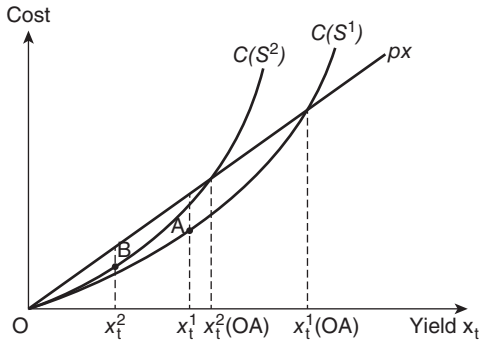


Figure 13.12 Cost function for harvesting various yields from a given stock.

yield curve $C(S^1)$ is parallel to the revenue line px at point A; the vertical distance from A to the revenue line, which is revenue minus cost, or profit, is the greatest at this yield. With a smaller stock, S^2 , the cost curve $C(S^2)$ operates, and the optimal yield is at x_t^2 , directly below the profit-maximizing cost at B. We will discuss the two yields labeled (OA) in subsection 13.2.4. If the resource price increases, line px will rise counterclockwise, and the profit-maximizing yields will rise. Alternatively, if the cost function for a given population size increases, the cost functions will twist counterclockwise and yields will fall. Connecting the profit-maximizing points in this diagram, we can transfer them to the growth curve as the yield curves of Figure 13.11(a).

We can develop the growth-yield diagram quite explicitly from the production function for yield and the growth/harvest relationship.⁷ Give a Cobb–Douglas form to the yield function to get $x_t = AS_t^\alpha e_t^\beta$, in which $A > 0$ and $1 \geq \alpha, \beta > 0$. Invert this relationship (that is, solve it) for effort as a function of stock and yield as shown above to get $e_t = c(S_t, x_t) = BS_t^{-\alpha/\beta} x_t^{1/\beta}$, which is the effort cost of yield. The market price is p (which is marginal revenue to a competitive producer), to which the owner of the resource will set his marginal harvesting cost: $p = w\beta^{-1}BS_t^{-\alpha/\beta} x_t^{(1-\beta)/\beta}$, where w is the unit cost of effort (the wage for labor or the rental on boats, or a combination thereof). Now bring back the net growth relationship $\Delta S_t/\Delta t \equiv \dot{S}_t = G(S_t) - x_t$ and substitute the solution for x_t in the $MR = MC$ relationship to get $\dot{S}_t = G(S_t) - p/w^{1/\theta}(\beta/B)^{1/\theta} S_t^{\alpha/\beta\theta}$, which is a

difference between two functions of the resource stock (simplifying the exponents); this expression simply says that the net growth of the stock is equal to its natural increase less the amount harvested. Figure 13.13 shows the two curves, from which the stationary solution for harvest rate x_t , if one exists, can be found and the difference between a sustainable yield and extinction related to the product price-input price ratio, p/w . In panel (a), the ratio p/w is relatively small, and a stable, sustainable harvest rate is consistent with a long-run population size of S^e . As long as the initial resource stock is not smaller than S_e , a larger growth than harvest will drive the population to S^e ; if it is below S_e growth is below harvesting and the population will die out. In panel (b), the yield curve for a higher p/w ratio never intersects the growth curve and no harvest rate is consistent with a sustainable population; the resource population will be harvested to extinction. Thus, more desirable renewable resources that are very easy to catch or harvest are likely to be collected to extinction.⁸

13.2.3 The theory of optimal use

Against this background of biological growth and the interaction of human predation on it, an agent owning a renewable resource will want to obtain the maximum possible value for it, now and in the future, whether the harvesting plan that will deliver that exhausts the resource (extincts a population or species) or is sustainable. The intertemporal maximization problem is much the same as in the case of the exhaustible resource: $\max_{\{x_t\}} \sum_0^T [p_t x_t - c(S_t, x_t)] / (1+r)^t$ subject to $\dot{S}_t = G(S_t) - x_t$. The first-order condition for the harvest rate is the same as in the case of the exhaustible resource: $p_t = \lambda_t + c_x$, that is, the market price exceeds the marginal harvesting cost by the royalty. The rate of change of the royalty over time has an additional term compared to that for the exhaustible resource case, an extra dividend accounting for the fact that waiting cost is reduced by the ability of the stock to grow: $\Delta \lambda_t / \Delta t \equiv \dot{\lambda}_t = r\lambda_t + c_S - G_S \lambda_t$. However for a sustainable harvest (a stationary state of population), both $\dot{\lambda}_t$ and $\dot{S}_t$ must equal zero, which implies that the stationary relationship between the market price and the royalty must be $p_t = \lambda_t(1 + G_S - r)$. Additionally, with

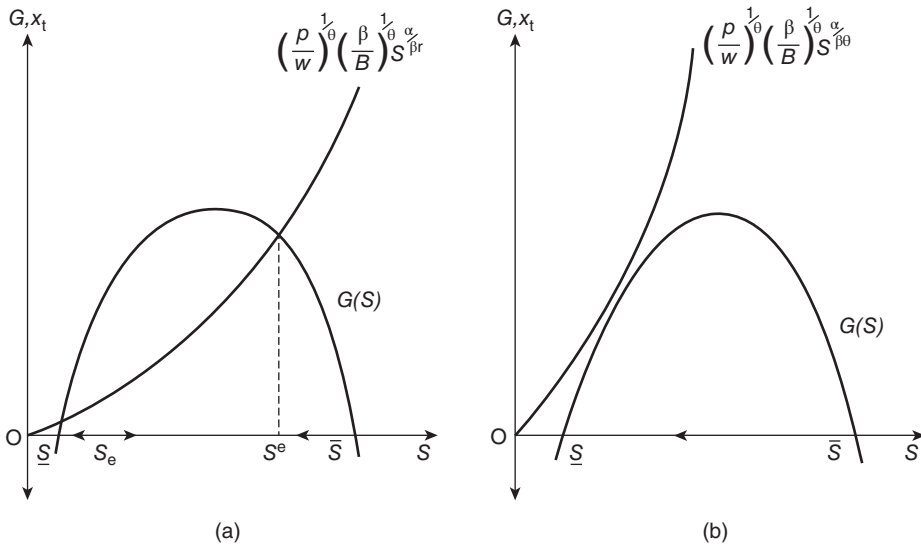


Figure 13.13 (a) Stationary harvest rate: population may be sustained or driven to extinction. Adapted from Dasgupta and Heal 1979, 124, Diagram 5.4. © 1980 Cambridge University Press. (b) Stationary harvest rate: no harvest rate consistent with sustainable population. Adapted from Dasgupta and Heal 1979, 124, Diagram 5.5. © 1980 Cambridge University Press.

unchanged demand, the market price of the harvested resource must remain constant over time as well: $\dot{p}_t = 0$.

13.2.4 Open access and the fishery

The classic example of the open-access renewable resource is the fishery, although grazing, hunting, and groundwater exploitation have the same structure. A definition of a fishery may be a useful beginning (Anderson 1977, 22). It includes a stock or stocks of fish and the enterprises that have the potential of exploiting them, which points to the importance of currently inactive fishermen, those waiting for the right opportunity, as well as those actively fishing. The biological variables are the size of the fish stock and its growth rate. On the human side, the variables are total fishing effort, catch, costs and revenues. Fishing effort can be allocated across the number of boats, their ability to catch, their spatial distribution, time spent fishing, the skill of the crews, and various types of equipment that may be subject to technical improvement such as nets and hooks. We will concentrate on numbers

of boats as the principal measure of fishing effort, but consideration of each of the others can add richness to the analysis of fishing in a particular place and time.

We can illustrate the structure of the problem facing users by comparing the production opportunities confronting a user of an open-access resource with the production function that we have already shown for the owned resource. With open-access resources the distinction between what the individual user sees and what the entire collectivity of users – the industry – sees is important. From the perspective of an individual boat owner, catch is a function of the fish stock and the number of other boats fishing, V_t : $x_t = f(S_t, V_t)$, $f_S \geq 0$, $f_V \leq 0$. A larger fish stock is likely to raise an individual fisherman's catch, other things, such as the number of other fishermen in the area, remaining the same; but a larger number of other boats competing for the fixed stock of fish will reduce his catch, and the additional boats also will reduce the sustained yield. The individual fisherman has control of neither of the variables in this production function, and this specification does not distinguish

variable hours spent fishing. From the perspective of the individual fisherman, the presence of the other boats is a pure externality. Similarly, depletion of the fish stock by the rest of the industry is an externality.

From the industry perspective, more boats catch more fish altogether, $X_t = \sum_{i=1}^V x_i$: $X_t = Vf(S_t, V_t) = F(S_t, V_t)$, $F_V > 0$, $F_{VV} \leq 0$, $F_{SV} \geq 0$. $F_V > 0$ means that more boats will catch more fish in total, while $F_{VV} \leq 0$ means that additional boats can be expected to add proportionally fewer fish caught, although over some range (the first few boats) the depressing effect on total catch may be nil. This is a crowding effect. $F_{SV} \geq 0$ means that the interactive effect of more boats and a larger stock of fish may (but not necessarily will) yield a larger catch.

Using the relationships between population, effort and catch in Figure 13.11, we can derive a relationship between revenue and effort that has the same quadratic shape as stock growth and sustainable yield, shown in the upper part of Figure 13.14. The figure applies to the industry as a whole rather than to a specific individual, who generally would be unable to affect sustainable yield in an entire fish population – at least not using ancient fishing technologies. While the total revenue curve is quadratic, paralleling sustainable yield, the total cost is a straight line, price times catch (assuming price is unaffected by the local industry's catch; if it were, this line would fall off to the right). Effort level e^1 is the profit maximizing level of industry activity, the slope of curve TR being parallel to TC at that point – the point of maximum economic yield. However, the fact that no one owns the fish leaves fishermen free to harvest them without payment for their scarcity value – no one can enforce claims to the royalty value of the fish. The total revenue they derive from selling the fish can go to paying for their time and the maintenance on the boat, leaving total revenue equal to total cost and consequently average revenue equal to average cost. As you can see from the lower panel of Figure 13.14, this leaves marginal revenue negative, which is also apparent from the upper panel: the total revenue derived from effort level e^2 could have been derived with far less effort, e^3 . Without some agency stepping in to enforce

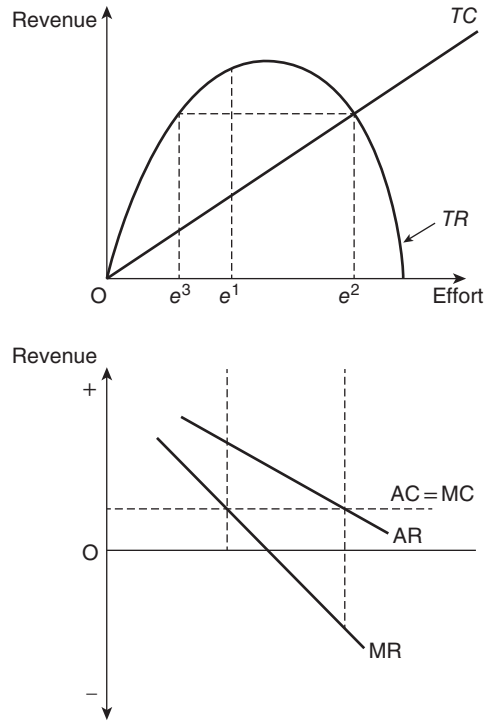


Figure 13.14 Relation between revenue and harvesting effort.

payment of the royalty value of the fish, the fishing industry competes itself into a region of effectively negative production.

The open-access characteristic yields some counterintuitive consequences of changes in technologies and costs. The following examples are from Anderson (1977, 48–52). First consider the response to changes in input costs to fishing – say the rental on vessels fell, or the cost of fishhooks dropped, or the opportunity cost of labor time decreased. Figure 13.15 shows a series of cost decreases, from yield cost curve c^1 through c^3 . The fishery represented by the relationships underlying curve TR will not be exploited with cost regime c^1 . A reduction in cost from c^1 to c^2 would permit exploitation up to effort level e^2 (effort level e^1 is zero). Suppose costs continue to fall to c^3 , and the industry expands its efforts to the point where total costs equal total revenue at e^3 . This move is counterproductive because the expansion reduces the fish population so

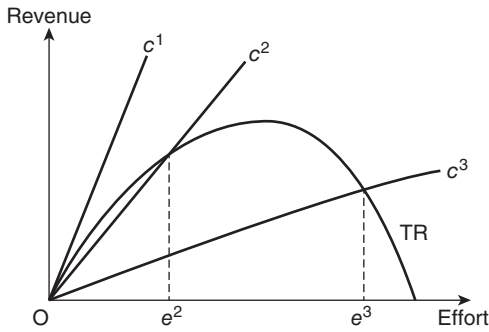


Figure 13.15 Effect of cost changes on harvesting effort. Adapted from Anderson 1977, 48, Figure 2.6. By permission of The Blackburn Press, Caldwell, New Jersey, United States.

much that sustainable yield is lower than it was with the lower effort, but the individuals in the fishery are unable to prevent the additional use. Cost reductions in providing fishing effort never benefit the fishermen in the long run because any profit generated encourages entry of additional fishermen – or extended hours of current fishermen – until the profit is entirely eliminated. The cost decreases we’ve considered between c^2 and c^3 actually reduce long-run yield. You’ll notice, however, that cost reductions between c^1 and c^2 (and those between c^2 and the maximum point on TR) will increase total revenue by driving the population into a more productive size, but the long-run effects for the fishermen are the same: profits are completely eroded by the open access.

For a technological improvement, suppose that superior nets are introduced that let fishermen catch the same number of fish in fewer days. Archaeological evidence on some of these technologies actually may be available in the form of surviving equipment – net sinkers, fishhooks, tridents – and depictions on pottery and later on mosaics. This shifts the revenue curve leftward to TR_1 in Figure 13.16 because each unit of effort with the new equipment is more effective than before. Effort falls from e^1 to e^{1*} . Catching this many fish – earning the same revenue – would have taken effort e^3 with the old nets, but even with the lower “nominal” effort the fisherman with their new nets depress the fish population to a size that offers a smaller sustained yield. One benefit to the individuals involved in this fishing,

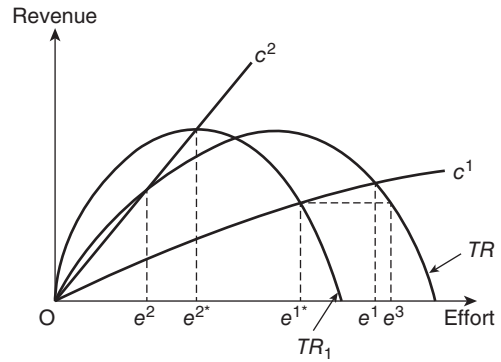


Figure 13.16 Effect of technological change on harvesting effort. Adapted from Anderson 1977, 49, Figure 2.7. By permission of The Blackburn Press, Caldwell, New Jersey, United States.

however, is that they may devote more time to other nonfishing lines of production where they do expand real output.

On the other hand, if the effort cost curve is c^2 , these new nets will raise effort from e^2 to e^{2*} . The greater effective effort still reduces the fish population, but in this case pushes it into a more productive size range that offers a larger sustained catch. Technological improvements in an open-access fishery always increase effective effort, but if the fishery is operating beyond the maximum sustained yield (to the right of the maximum point on curve TR in Figure 13.16), fishermen will decrease nominal effort and cost. If it is operating in a region of increasing sustained yield (to the left of the maximum of TR), fishermen will increase their nominal effort – their actual hours in the boats – and cost.

Next consider the consequences of price changes, or alternative prices in another interpretation. Higher fish prices leave the two zero points of the revenue curve unchanged because those points depend only on the physical relationship between catch and regeneration. Higher prices expand the height of the parabola representing revenue, as in the shift from TR_0 to TR_1 in Figure 13.17. With prices yielding revenue curve TR_0 , this fishery is not exploitable with either technology represented by effort cost curves c^1 or c^2 . With the prices giving revenue curve TR_1 , the fishery is exploitable, and effort e^1 will be applied to it. If prices were higher – or if they increased – and generated revenue curve TR_2 , effort would increase to e^{1*} with technology 1,

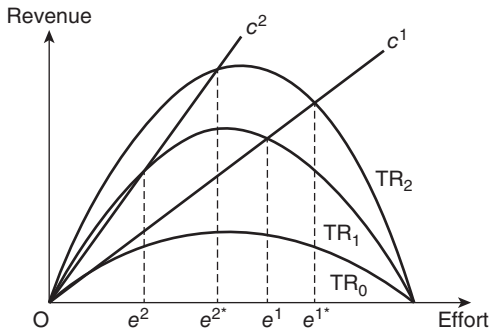


Figure 13.17 Effect of product price changes on harvesting effort. Adapted from Anderson 1977, 51, Figure 2.8. By permission of The Blackburn Press, Caldwell, New Jersey, United States.

and revenue would be higher but the actual catch would be smaller than before because the fishermen had already pushed the fish population past its maximum sustainable yield level. Look back at either of the previous revenue curves, TR_0 or TR_1 , and you'll see immediately that the constant-price revenue is smaller at e^1 than at e^{1*} , meaning that the increase in total revenue comes from a smaller number of higher price fish. If the catch technology were described by cost curve c^2 , fishermen would experience higher prices *and* catch more fish because their increased effort, from e^2 to e^{2*} , pushed the fish population into a smaller but more productive range of stock size.

13.3 Resource Scarcity

Resource scarcity is intuitively a physical concept: how much of some natural resource remains to be used? However, the reserves known or estimated to exist at a point in time are subject to considerable uncertainty and ignorance and may be updated radically with little warning. Additionally, the known reserves at any time can be augmented by deliberate exploration. Between these two sources of flexibility in the definition of reserves, we are left with little of solid usefulness in yielding information about scarcity.

A price concept captures the relative scarcity reflected by how much of it is demanded in comparison to how much is known to exist and how costly it is to extract. Current prices, in contrast

to futures prices (few such markets probably ever existed in antiquity) or forecasts of future prices (which had to be made in antiquity to decide about investments in mine development) possess the advantage of observability but they are subject to periodic revision following discoveries of new reserves as well as to the effects of technological advances in both extraction and use. Examinations of long time series of resource prices have tended to show little increase in scarcity in the past century for most of them. Royalties – the measure of the value of holding a unit of a resource stock rather than extracting and consuming it currently – come close to what is desirable in a measure of scarcity but they are not readily observable.

However, remember the close relationship between exploration costs and royalties when we included prospecting for new reserves in the optimal depletion or use program. The amount that extractive enterprises can afford to spend in finding a new unit of reserves is equal (or at least quite close) to the value of having an additional unit of reserves – it's only a difference of whether it's just sitting there and you just have to hold on to it rather than mine it or whether you have to go out and find it. So the marginal cost of exploration would make an excellent substitute for direct information on royalties. Still, observability can be a problem for marginal exploration cost, but the average cost per unit of new reserves discovered may be a satisfactory approximation. Direct evidence for antiquity on exploration costs, divided by the quantity of additional reserves discovered, sounds like a Christmas wish, but retaining in one's mind that the amount spent on exploration, even in antiquity, is some indicator of how much the new reserves would have been worth when discovered could help extract some information out of the scanty evidence on the subject that might come available here and there.

13.4 The Ancient Mining-Forestry Complex

The great nonrenewable resource exploitation industries of the ancient Mediterranean, the smelting of the seven principal metals of antiquity – copper, tin, arsenic, lead, iron, gold,

and silver – operated closely with the exploitation of one of the principal renewable resources, forests. Pyrotechnology during the period was not particularly efficient, although it did improve, in terms of wood-to-heat ratios, probably because of scarcity of appropriate wood. Prior to the introduction of metal axes and saws, many of the forests probably had regeneration rates in excess of the extractive capacity of the people exploiting them. The introduction of metal tools raised the feasible rates of exploitation of forests, probably beyond natural growth rates for many species (Wertime 1982).

Ores mined from the available mineral deposits were smelted with charcoal from local – and probably not so local, at times – forests. Choice of a depletion rate for a mine, if it ignored the sustainable yield of forests within accessible transportation distance, would have run the risk of deforestation prior to exhaustion of the mineral deposit. Archaeological evidence of transportation of ores to smelting sites overseas might suggest local deforestation in the mining region and cheaper transportability of ores than of wood in sufficient quantities.

Wertime (1982) has suggested that the lower heat demanded by iron than by copper played a principal role in the emergence of iron smelting technology and the eventual supplanting of bronze with iron in a wide variety of uses, probably beginning in the thirteenth and twelfth centuries B.C.E. Metallurgical investigations of slags throughout the Aegean indicate that smelters were conscious of scarcity of fuel and used options to reduce heat demands in smelting a variety of ores. The forests appear to have been a case of the “high p/w ratio” in the harvest function, relative to the position of the resource growth curve, depicted in Figure 13.13(b), probably in conjunction with having been an open-access resource.

Metal smelting was not the only important source of demand for charcoal-fired heat in the ancient Mediterranean. Pottery, glass products, limestone for cement, and plasters all possessed derived demands for trees for fuel, and Wertime discusses how the technologies in various industries probably led to the discovery of new products in other industries, expanding the demand for fuel but improving heat ratios from

charcoal only marginally. The nonrenewable resources used in these industries may occasionally have been depleted in the cases of particular stone quarries or pottery-quality clay beds, but shrinking forestry resources raising the price of fuel probably more commonly was responsible for substitutions of resource sources and industrial outputs at various locations.

13.5 Suggestions for Using the Material of this Chapter

The exhaustible resources of antiquity are primarily mines, quarries, and clay pits. The chapter introduced the concept of economic exhaustibility to supplement that of physical exhaustion – simply mining something out. Economic exhaustibility tells us that when cheaper sources of an exhaustible resource are found, higher cost sources won’t be able to compete and will shut down – or at least sharply curtail their rates of output. Current ore contents of abandoned mines may reflect the existence of ancient competitors as well as the prowess of the ancient technologies.

The renewable resources would have included forests, fish, game, and I would include soil. Pollen analyses and soil studies have suggested periods of deforestation and other vegetative change, and studies of animal bones can give some idea of recent elimination of species, whether through over-hunting or deliberate suppression of species considered competitors to humans in agriculture and stock raising. Societies’ actions to preserve soil, either physically through suppression of erosion with terracing or through cultivation practices intended to restore nutrient content, can be found both in the archaeological record and in textual evidence. The science and technology of salt accumulation from irrigation apparently was incompletely understood in antiquity, however, and many hundreds of square miles of irrigated farmland were lost to salt pans in Mesopotamia. That said, given the options for using the land – possibly not at all without irrigation, and maybe for five or ten years with irrigation – the choice may have been the ancient equivalent of the no-brainer. Reclamation of swampy land could be added to the types of soil preservation practiced in antiquity, possibly such

as in the Fayoum in Egypt during Ptolemaic times or the draining of Lake Copais in Boeotia during the Greek Archaic period. Ancient deforestation

has been noted in Greece (Bottema 1982) and Crete (for example, Rackham and Moody 1996, Chapter 11).

References

- Anderson, Lee G. 1977. *The Economics of Fisheries Management*. Baltimore MD: Johns Hopkins University Press.
- Bottema, S. 1982. "Palynological Investigations in Greece with Special Reference to Pollen as an Indicator of Human Activities." *Palaeohistoria* 24: 257–289.
- Conrad, Jon M., and Colin W. Clark. 1987. *Natural Resource Economics: Notes and Problems*. Cambridge: Cambridge University Press.
- Dasgupta, Partha. 1982. *The Control of Resources*. Cambridge MA: Harvard University Press.
- Dasgupta, P.S., and G.M. Heal. 1979. *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.
- Fisher, Anthony C. 1981. *Resource and Environmental Economics*. Cambridge: Cambridge University Press.
- Hotelling, Harold. 1931. "The Economics of Exhaustible Resources." *Journal of Political Economy* 39: 137–175.
- Pindyck, Robert S. 1978. "The Optimal Exploration and Production of Nonrenewable Resources." *Journal of Political Economy* 86: 841–861.
- Rackham, Oliver, and Jennifer Moody. 1996. *The Making of the Cretan Landscape*. Manchester: Manchester University Press.
- Smith, Vernon L. 1975. "The Primitive Hunter Culture, Pleistocene Extinction, and the Rise of Agriculture." *Journal of Political Economy* 83: 727–755.
- Waldbaum, Jane C. 1980. "The First Archaeological Appearance of Iron and the Transition to the Iron Age." In *The Coming of the Age of Iron*, edited by Theodore A. Wertheim and James Muhly. New Haven CT: Yale University Press, pp. 69–98.
- Wertheim, Theodore A. 1982. "Cypriot Metallurgy against the Backdrop of Mediterranean Pyrotechnology: Energy Reconsidered." In *Early Metallurgy in Cyprus, 4000–500 B.C.*, edited by James D. Muhly, Robert Maddin, and Vassos Karageorghis. Nicosia: Department of Antiquities, Cyprus, pp. 351–361.

Suggested Readings

- Conrad, Jon M. 2010. *Resource Economics*, 2nd edn. Cambridge: Cambridge University Press.
- Dasgupta, P.S., and G.M. Heal. 1979. *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.

Notes

- 1 The classic analysis remains Hotelling (1931).
- 2 Why not an extraction rate for the terminal period? We'll see that extraction is equivalent to the difference between the stocks at the beginning of adjoining periods, and S_T is the amount of the resource left in the ground when the stock is considered economically exhausted.
- 3 I could have specified an exploration production function in which more exploration effort finds more new reserves and invert it to get a cost function like c^2 . Exploration cost functions are commonly made dependent on cumulative previous discoveries, with current costs of discovering another unit of the resource higher the larger are cumulative previous discoveries – call them D ($c_D < 0$) – implying a depletion effect in exploration as well as in mining of a fully fixed stock.
- 4 There's actually a little sleight of hand here, because there are separate Lagrange multipliers for each constraint, and the additional term for exploration costs will bring with it an additional constraint on resources put into exploration efforts. The Lagrange multiplier for that constraint is distinct from the one for extraction, whose first-order condition yields the relationship that price equals marginal extraction cost plus royalty. The multiplier on the new constraint, call it μ , is the shadow price of an additional unit of cumulative discoveries, which captures the effect of this unit on future marginal discovery costs. As we noted in footnote 3, as cumulative discoveries are expected to raise future

discovery costs, this effect would be negative. Nonetheless, it could be positive, and we can suppose the effect to be small or zero, which gives us the result that the marginal discovery cost of new reserves equals the royalty, or the shadow value of another unit of the ore in the ground. In the text, I've shortcut the second multiplier by assuming its value to be zero for simplicity. If it is nonzero, the royalty would be somewhat above the marginal discovery cost. See Pindyck (1978).

- 5 Useful treatments of monopoly are given in Fisher (1981, 37–44) and Dasgupta and Heal (1979, 323–334).
- 6 With fish it is reasonable to track individuals, although for more slowly growing species

accounting for growth of existing individuals as well as births of new ones is important. In the case of trees, the population of a managed stand is fixed, and the biomass grows in each period. Both population and biomass growth will follow such logistic growth.

- 7 For the application of this analysis to open-access fishing, see Dasgupta and Heal (1979, 122–125) and Dasgupta (1982, 125–129).
- 8 Smith (1975) finds this relationship capable of replicating the extinction of large mammals such as the woolly mammoth at the end of the Pleistocene as well as the virtual extinction of bison in the American West in the late nineteenth century.

Growth

14.1 Introduction

14.1.1 *Economic growth: delimiting the scope*

Economic growth is a candle to economic-historian moths. Conceived as the story of the progress of humanity from barely surviving from season to season to a worldwide epidemic of obesity – granted, interspersed with various people starving to death – historians of many periods, as well as scholars from various social science disciplines, periodically revisit the subject in intensities that have the appearance of a bandwagon effect. Pushed to the background for a generation or so, growth reappears now and then as “the” subject to explain or discover or measure. The origins, causes, locations, and so on of the eighteenth-nineteenth (or occasionally pushed back to the seventeenth) century C.E. Industrial Revolution have occupied economic historians of those periods fairly intensively since the 1970s, and their research has served recently as inspiration to scholars of ancient societies, at least those of the Mediterranean-Aegean region, if not necessarily spreading to those who study ancient Egypt and the ancient Near East. Insights of varying confidence have been obtained from research since 2004 on growth in antiquity. To make a contribution to continued research on

this topic, the purview of this chapter must be delineated clearly and what it hopes to achieve and to avoid specified.

First, this chapter uses the term “economic growth” to mean increase in production per worker or consumption per worker (any differences between the two being attributable to institutional arrangements for distributing the product). Aggregate growth, a combination of production per worker and the number of workers, is an interesting topic, but involves as much population growth as growth in individual productivity, the latter of which is the subject of economic growth theory. There are interactions between total population size and individual productivity growth, but they are secondary although not insignificant. For example, a larger number of brains at work at the same time, given any probability of one of them thinking of something useful, means a higher probability of discovering something that will be useful to human productivity – that is, innovation, if we take all the steps past discovery or invention. Then there is Smith’s dictum about the division of labor being determined by the extent of the market, which means that the larger a market is – that is, the more people working in it, the more specialized individual producers can become, raising their productivity through concentration on one set of

operations. True enough, but taking one issue at a time, the chapter focuses primarily on causes of economic growth other than larger market sizes.

Second, the chapter is intended to offer guides to thinking about and studying economic growth in ancient societies rather than to enter the empirical fray of whether growth actually took place in certain regions over certain periods and if so, what might have contributed to it and what might have contributed to its cessation. A number of possible indicators have been used as evidence of growth, for example: skeletal evidence of nutrition and general health, including diseases associated with population crowding; artistic advances such as those in Greek vases between the Sub-Mycenaean and Late Geometric Periods; shipwrecks; and dwelling sizes.¹ Those subjects are, strictly speaking, outside the purview of the chapter's goals, which is exposition of the theory of economic growth and closely related topics in a manner accessible for scholars of antiquity. However, when understanding could be enhanced by linking elements of theory to the context of ancient evidence and current thinking on it, examples are brought in. Actual changes in production per worker, the subject of this chapter, may be more difficult to detect directly in evidence from antiquity than the indicators used to date. Additionally, the correspondences between ancient productivity changes and the indicators that have been used are sometimes more intricate than assessments have dealt with.

14.1.2 Growth in antiquity: is there anything to explain?

Prior to the advent of the Industrial Revolution, beginning in Britain in the eighteenth century, the record of personal income levels from previous periods extending back to antiquity can seem dismal. From the low starting point of the typical Western European peasant of the Enlightenment Period, it is difficult to think of people having substantially less and still surviving, although shrinkage in physical stature can be an accommodation to lesser nutritional availability. The entire period from the later

years of the Roman Empire in the West until the mid- to late eighteenth century, and possibly even later, is one of virtually no increase in production or consumption per capita.² What could there be to explain in antiquity with growth theory?

Just as different regions of the world existed at different levels of sophistication at any one time in antiquity, those levels of sophistication and per capita production (consumption, income, wellbeing) were reached from lower levels. The paths in between were periods of economic growth, even if they were not sustained for all time, and may even have regressed. Many of the early civilizations reached their pinnacles and declined for one reason or another: the Sumerian and Akkadian, the Old Babylonian, earlier and later Assyria, the second and first millennium city-states of the Levantine littoral, Minoan Crete, Mycenaean Greece, Classical Athens and some of the other Greek city-states of the period,³ and of course Rome.⁴ Some, such as Egypt, experienced several millennia of rises and declines of fortunes. We can think of these successive periods as the results of wars, weak government, environmental stress, or such anthropomorphisms as societal aging and decline. Each of these viewpoints may offer insights into the processes, if not always specific events, involved in these enormously complicated histories or prehistories. Concepts from the theory of economic growth and from the somewhat more sprawling body of thought on economic development may help clarify thinking at certain points. The absence of growth – or zero growth rates in per capita output and income – is as important a phenomenon to explain as periods of growth, even if that growth was not sustained.

14.2 Essential Concepts

14.2.1 Production functions again

The central core of the study of economic growth is the production function, the relationship between the quantities of inputs and the quantity of output. Growth is, in fact, nothing more than

producing larger quantities of outputs from given quantities of inputs. The production-function paradigm is a useful reminder of the fundamental conservation law of economics: that nothing comes out of nothing. Changes in output, either total or per capita, derive from certain, well-known changes such that if we observe a change in either measure of output we know that some other events must have occurred. Correspondingly, we also know that if those events occurred, growth must have taken place. Total output will not increase without an increase in at least some input, not necessarily all inputs but at least one. Output per capita will increase necessarily when the capital-labor ratio increases. These two theoretical relationships give us some reliable places to look when we think that growth either did or should have occurred at some time and place.

14.2.2 *Technical change*

Every society has its understandings of how the world works. Much of this understanding is embodied in the technology of its capital equipment. Although technical change has been institutionalized and made systematic within the paradigm of modern science in the last several centuries, combinations of trial-and-error and fortuitous accidents raised the productive power of capital equipment over the centuries of antiquity, if slowly. Technical change can raise the productivity of capital, labor, materials, or any combination of those three inputs. In the production-function specification, technical change alters the effective quantity of one or more inputs and increases output per unit of input, changing the marginal products of inputs.

Even without technical advances in capital equipment, given time, people will figure out more effective ways to employ the resources at their disposal, from ways to pack jars in ships, to the best timing of planting seeds. Some of these improvements derive from organizational changes, with unchanged equipment, and it can be splitting hairs to avoid calling these changes technological advances simply because they occur in what would be called “soft technology” today.

Nonetheless, representing them in a production function can be difficult, and it is useful to keep the distinction between technical changes in capital and organizational changes.

14.2.3 *Growth versus development*

In the study of the transformation of entire economies from social and technological systems not substantially different from those of medieval times to predominantly market-organized, high-per-capita-income, relatively capital-intensive, and predominantly Western, societies that we call “modern” and industrialized, the distinction between “growth” and “development” has been found useful, if not always easy to define. “Growth” can be defined pretty clearly to mean an increase in per capita income or output. “Development” refers to the organizational and institutional characteristics of the society. In the second half of the twentieth century, “development” came to imply transformation from a predominantly rural and agricultural society and economy to one predominantly urban and industrial. Institutions emerged that supported more anonymous relationships among producers and between producers and consumers, provided general education and specialized training, supported greater geographical and occupational mobility, generally supported the extension of markets to a wider range of goods and services, and accomplished many other tasks that served the role of a social integumentary system.

Students of antiquity will recognize immediately that institutions existed that performed each of these tasks. Just as institutional change that can be documented supported the urbanization of Western Europe and North America in the nineteenth and twentieth centuries, institutions surely emerged to support the growth of the urban systems of antiquity, even if we cannot see many of them all that clearly today. We catch glimpses of the educational and training systems of antiquity through occasional tablets and papyri describing apprenticeships and the role of kinship in sustaining occupations. Differences in levels of urbanization and overall social sophistication existed among the major regions

of the Mediterranean. They must have been accompanied by institutions that supported the interactions required to sustain those differences. Extended periods of economic growth are likely to be accompanied by the institutional changes subsumed under the rubric of development, even if we have to loosen our concept of “development” beyond a definition of “becoming more like us,” as dominates the contemporary, policy-oriented discussions of development.

14.3 Neoclassical Growth Theory

I still recall the introduction one of my teachers gave to this subject: “There’s less to growth theory than meets the eye.” Despite the importance of the subject and its extensive mathematical development, this still strikes me as essentially true, possibly reflecting how little we still really know about the causes of economic growth. However, if we can grant a few causal kick-starts, neoclassical growth theory offers some useful models of what does and doesn’t happen. A number of models of economic growth have dead-ended during short half-lives after World War II, but the one that has proven the most robust, and recently served as the springboard for further development, derives from the neoclassical production function. Trevor Swan and Robert Solow published models in this format at virtually the same time in 1956, although Solow’s name, and his subsequent studies of the subject, gained name dominance as well as the Nobel Prize. The critical advance contributed by Solow and Swan was to take advantage of the factor substitution emphasized in the neoclassical production function, in contrast to fixed-coefficients approaches to growth theory that preceded their models. The fixed-coefficients approaches predicted that economies would encounter serious trouble if they ever got into situations in which their investment rates offered either more capital than was needed or less – they would spiral away from any equilibrium combination of employment and factor prices (wages and rents), in a fashion simply not observed. Despite some complaints

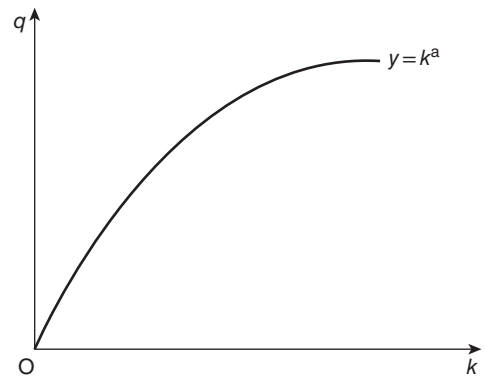


Figure 14.1 Per capita output as a function of the capital-labor ratio.

about the Solow-Swan model, and recent modifications to remedy those complaints, it remains the foundation of contemporary thinking about economic growth.

14.3.1 The Solow model

The Solow model is built around two relationships, a production function and a capital accumulation equation.⁵ Let a single production function represent an entire economy, which implies that the economy produces and consumes a single good: $Q = F(K, L)$, or using specifically the Cobb–Douglas form, $Q = K^\alpha L^{1-\alpha}$, where Q is output, K is capital, L is labor. Letting the output price be equal to 1, profit-maximizing firms solve the problem $\max_{\{K, L\}} F - rK - wL$, whose first-order conditions give $w = \Delta F / \Delta L = (1 - \alpha)Q/L$ and $r = \Delta Q / \Delta K = \alpha Y/K$. Then the sum of payments to factors of production equals the (value of) output: $wL + rK = Q$. Expressed in terms of output per worker, $q = f(k, 1)$ or simply $f(k)$, where $q \equiv Q/L$ is output per worker (or per capita since we can equally well suppose that all people are workers), and $k \equiv K/L$, or the capital / labor ratio. The Cobb–Douglas form of the production function is, then, $q = k^\alpha$. Figure 14.1 depicts the production function in per capita terms. This version of the production function tells the amount of output per person produced

for whatever capital stock per person exists in the economy.

The other relationship describes how capital accumulates: $\dot{K} = sQ - dK$, where $\dot{K} \equiv \Delta K / \Delta t$, where t represents time, so that $\dot{K}$ is the absolute growth (not the growth rate yet) of capital over a short time period; s is the savings rate out of consumption, and d is the depreciation rate. The first term, sQ , is gross investment and the second term, dK is the total amount of depreciation that occurs during production. The basic version of the Solow model assumes that workers (which amount to consumers, since we assume all people work) save a constant fraction s of their combined wage and rental income. With a closed economy (one that does not trade), saving equals investment, and the only use of investment is to accumulate capital. The fixed depreciation rate d means that a constant fraction of the capital stock depreciates (disappears, is used up) each period regardless how much output is produced, or equivalently, how hard the capital is worked.

To examine the path of output per person over time, we can rewrite the accumulation relationship in the capital-per-person form as $\dot{k} = sq - (n + d)k$.⁶ This expression shows that the change in capital per worker in each period is determined by three forces. Two show up clearly in the original accumulation expression: investment per worker, sq , increases k , while depreciation per worker, dk , reduces k . The new term in this version is nk , the reduction in k caused by population growth. Each period sees the arrival of nL new workers who were not present in the previous period; with no new investment (or depreciation) there would be less capital per worker available because of the increase in the labor force by exactly nk .⁷

Figure 14.2, known as the Solow diagram, graphs the two terms on the right-hand side of the per capita accumulation equation. The curved line, which looks very much like the production function in Figure 14.1 (because it is the per capita production function, just scaled down by the factor s),⁸ represents the gross investment in each period, while the straight line from the origin is the amount of new investment per person

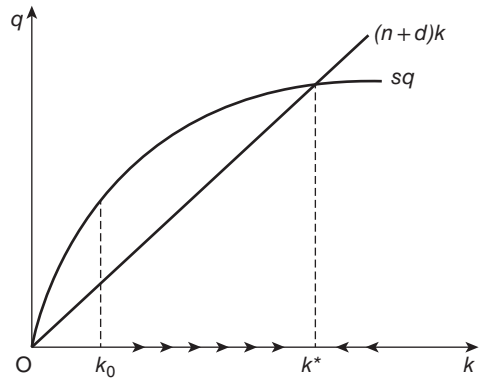


Figure 14.2 Changes in the capital-labor ratio.

required to keep the capital/labor ratio constant. The vertical difference between these two curves is the change in the amount of capital per worker. When this change is positive, the economy is increasing its capital per worker, which is called *capital deepening*. When the change is zero, but the total capital stock is increasing because population is growing, it is said that only *capital widening* is occurring. In Figure 14.2, if the economy were to start at k_0 , investment per worker exceeds the amount needed to keep the capital per worker constant, so capital deepening occurs – the capital/labor ratio, k , increases over time – until $k = k^*$, where $sq = (n + d)k$, so that $\dot{k} = 0$. At this point, capital per worker stays constant at what is called a *steady state*. If this economy had begun with more capital per worker than k^* , k would fall back to k^* . Growth rate $\dot{k}$ would be negative until the economy reached that capital / labor ratio.

We can use the Solow diagram to determine the steady-state capital stock per worker, total production per worker, and per capita consumption. Figure 14.3 shows the production function and the investment curve on the same diagram. The lower part of this diagram is just the Solow diagram, in which the intersection of the sq and $(n + d)k$ curves determines the capital/labor ratio, k^* . The production function lies above the investment curve, and extending the line from the optimal capital/labor ratio up to the production function, q , tells us the total output

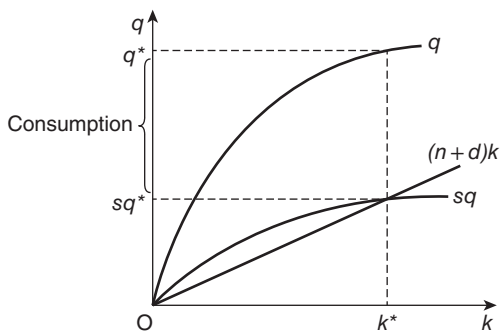


Figure 14.3 Steady-state capital-labor ratio.

per capita generated by this input ratio. Reading across to the vertical axis, q^* is the optimal output per capita. Consumption is production minus investment, or $q^* - sq^*$, identified in Figure 14.3.

We can use the same diagram to explore several important changes. Figure 14.4 raises the investment rate (or saving rate) from s_0 to s_1 , rotating the investment curve upward, or counterclockwise around the origin, from s_0q to s_1q . The investment requirements curve remains unchanged, and the optimal capital stock per worker increases from k_0^* to k_1^* . We can see how the change occurs from the diagram. At the original k_0^* , once the saving rate rises to s_1 , more savings is generated than is required to maintain the capital/labor ratio at k_0^* , so capital deepening

begins, which increases the capital / labor ratio as far as k_1^* , where investment generated just equals the amount required to maintain the capital/labor ratio constant, given the population growth and depreciation. At k_1^* , the economy is richer than it was before: the projection of k_1^* onto the production function (not shown in Figure 14.4) yields a higher q^* .

An increase in the population growth rate (not simply in population) causes the investment requirements line, $(n+d)k$, to rotate counterclockwise around the origin while the savings curve remains unchanged. In Figure 14.5, with population growth rate n_0 , the optimal capital/labor ratio is k_0^* , but with a higher population growth rate, n_1 , the same investment rate will not generate enough new capital to maintain the capital/labor ratio at k_0^* . Consequently capital per worker falls to the new equilibrium level at k_1^* , where the economy produces less and is poorer.

The steady-state amount of capital per worker is determined by the condition that $\dot{k} = 0$. We can substitute this condition into the per capita accumulation equation to solve for the optimal capital/labor ratio and output: $k = [s/(n+d)]^{1/1-\alpha}$; and then substitute this solution into the per capita production function to solve for optimal output, $q = [s/(n+d)]^{\alpha/1-\alpha}$. These two solution expressions indicate that countries with higher savings (investment) rates will, other things being equal,

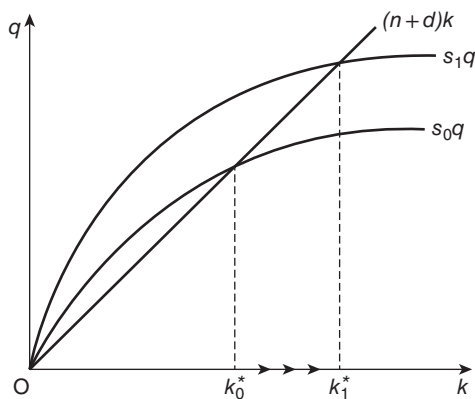


Figure 14.4 Effect of increase in the investment rate.

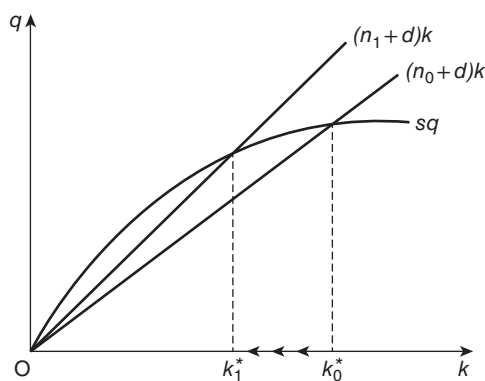


Figure 14.5 Effect of increase in the population growth rate.

be richer. Similarly, countries with higher population growth rates (not populations, but growth rates) will be poorer, also if other things are equal. What is the equilibrium per capita growth rate of output? Zero! q and k will be constant even in the face of a growing population as long as the saving rate remains constant. If the investment rate rises, the growth rate of output will rise for a short time as the capital / labor ratio moves to its new, higher equilibrium level, but once it reaches that value of k , the growth rates of both k and q will return to zero. As the capital / labor ratio increases, the marginal productivity of capital (the rental rate) declines to the point where it is unprofitable to use any larger quantity of it per worker.

14.3.2 Technology and growth in the Solow model

The Solow model isolates the key to economic growth – growth in per capita output or consumption – as technological change. There are several ways to represent technological change – not its causes but its consequences in production: the innovations can shift the entire production function upward, affecting the productivity of both labor and capital equally (Hicks-neutral technical progress); or it can raise separately the productivity of labor (labor-augmenting, or Harrod-neutral, technical change) or capital (capital-augmenting, or Solow-neutral, technical change). In the Cobb–Douglas specification of the production function, Hicks-neutral technical change is represented by $Q = AK^\alpha L^{1-\alpha}$, where A is an index of technology. Technological progress occurs when the index A increases over time, represented notationally by $\dot{A}/A = \dot{a}$ as the rate of technical change. Harrod-neutral technical change, the case Solow examined in his initial study, is represented in the Cobb–Douglas specification as $Q = K^\alpha (AL)^{1-\alpha}$.

To see how this specification of technology affects growth, express the production function in rate-of-change form: $\dot{q}/q = \alpha \dot{k}/k + (1-\alpha)\dot{A}/A$. From the accumulation equation, $\dot{K}/K = sQ/K - d$, a constant growth rate of capital requires a constant Q/K , which implies the constant per capita ratio, q/k ; q/k can stay constant

only if q and k grow at the same rate. If capital, output, consumption, and population are all growing at the same rate, we say that the economy is on a balanced growth path. In the Solow model without technical change, population could grow but the long-run growth in output and consumption per capita and capital per worker was zero because the rate of technical change was zero.

To examine the Solow model with technical change, we will redefine a few variables. The capital / augmented-labor ratio will be $c = k/A$ and the output per capita will be $y = q/A$. We can think of c as the capital per effective unit of labor, where the technical progress increases the effective quantity of labor that each worker offers. To get to the capital accumulation equation of this model, begin with the decomposition of the capital/effective-labor ratio: $\dot{c}/c = \dot{K}/K - \dot{A}/A - \dot{L}/L$. Combining this with the basic accumulation equation, with some substitutions, we get $\dot{c} = sy - (n + a + d)c$, graphed in Figure 14.6, which looks all the world like Figure 14.2.

We can find the steady-state capital / adjusted-labor ratio by substituting the production function into the per capita accumulation equation and setting that growth rate to zero, just as in the no-technical-change case: $c^* = [s/(n + a + d)]^{1/(1-\alpha)}$. Substituting this expression into the production function gives the steady-state output per adjusted worker: $y^* = [s/(n + a + d)]^{\alpha/(1-\alpha)}$.

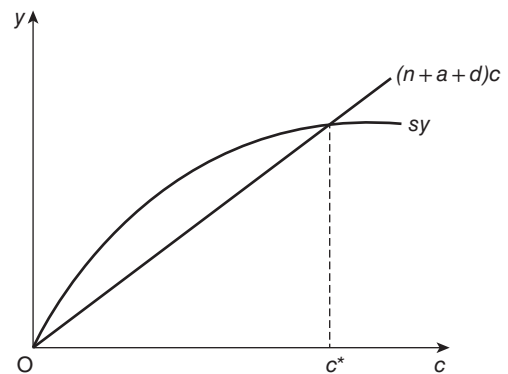


Figure 14.6 Effect of technical change on the capital-labor ratio.

To get back to output per worker, we extract the technology index from y^* to get $q^*(t) = A(t)[s/(n + a + d)]^{\alpha/(1-\alpha)}$, which indicates that output per capita at any time t depends on the level of technology at that time.

Just as in the case of no technical change, an increase in the savings rate (which is equivalent to an increase in the investment rate) has no long-run effect on the economic growth rate. Figure 14.7 uses the Solow diagram to show an increase in the savings rate when technological change is present, with identical results to the case of no technical change shown in Figure 14.4, despite the fact that a positive rate of technical progress is incorporated in the investment requirements line. However, let's follow the changes in the growth rates of the capital / labor ratio and output per capita to look more closely at the temporary jump in growth rates. From the initial, equilibrium capital / labor rate c_0^* in Figure 14.7, the increase in the saving rate gives an immediate boost to the capital / labor ratio as the economy moves from c_0^* to c_1^* because capital grows more quickly than labor, and the growth rate of output jumps correspondingly. As the economy approaches c_1^* , capital is added to the existing stock per worker more slowly, and the growth rate of output, while still positive (actually, positive and above the rate of technical change), also slows toward the rate of technical change. This initial jump and gradual return to the steady-state growth rate (equal

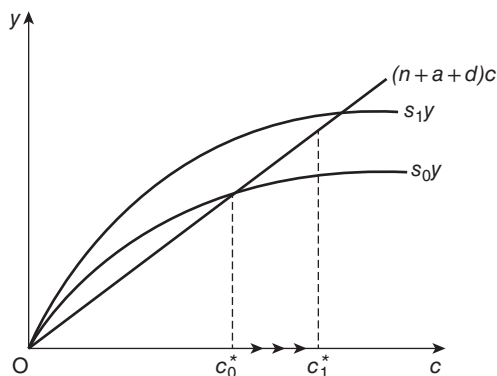


Figure 14.7 Increase in the savings rate.

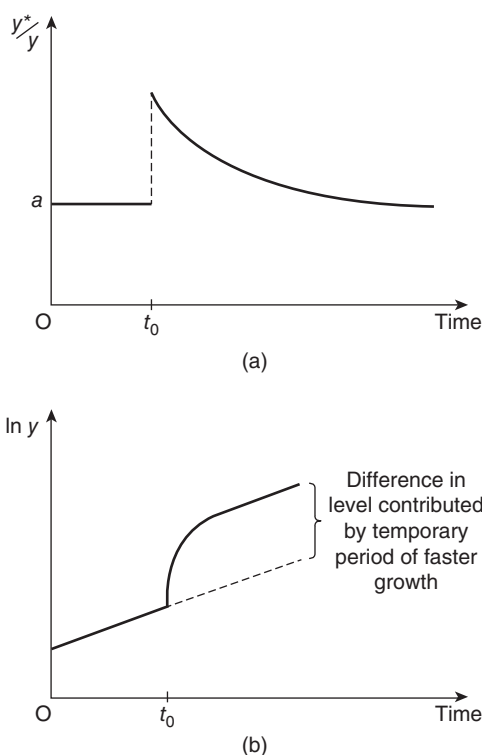


Figure 14.8 (a) Time path of output growth with increase in the savings rate. (b) Growth of output with temporary increase in the savings rate.

to the rate of technical progress) is shown in Figure 14.8(a). Figure 14.8(b) shows the time path of the level of output (in logarithms, since the exponential growth rates would present this graph with a rapidly – exponentially – increasing level): prior to time t_0 output (in logarithmic form) is increasing at a steady rate; when the saving / investment rate takes a jump at t_0 output begins growing considerably faster but fairly quickly slows down to its original growth rate, which was governed entirely by the rate of technical change.

The only action that will permit sustained economic growth is sustained technological progress. Where technical progress comes from – why it occurs – is outside the scope of the original Solow model but has been the subject of an immense amount of research in a number of

disciplines since the 1960s. Since the mid-1990s some methods have been developed for making the technical change in the Solow model endogenous – that is, emerging out of profit opportunities contained in the growth model itself rather than being specified exogenously as a given rate of technical change. We introduce these newer concepts in the following subsection, although with their technical complexity, and their heavy dependence on relatively recent institutional developments, we will not pursue them at length.

14.3.3 *Endogenizing technical change*

The key to obtaining sustained economic growth is to discover some way or ways that technical change can occur endogenously within the growth model. This is not a matter of tinkering with toys but rather one of discovering what mechanisms might operate in the world, and will continue over time, to raise the level of technology used in production. We can divide the categories of mechanisms into two principal types: accidental and purposive.⁹ The accidental discovery has a strong random component that does not make it a good candidate for generating continuous improvements in technology that would sustain economic growth, although it can make for good stories. In the purposive mechanisms, individuals recognize that there is profit to be made from improving technologies and deliberately strive to make those profits, in fact to maximize them. A little more taxonomy right now might not hurt, although taxonomies are a relatively low order of economic information. Within the purposive mechanisms it is useful for our purposes here to distinguish between the upgrading of labor – the accumulation of human capital – and technological improvement as a distinct productive activity in addition to production of consumption and capital goods. The second of these subcategories has generated a number of models of organized research and development which itself depends on an effective organization of science, not simply the sporadic, inquiring minds of genius that have dotted the records of antiquity. Despite the interest and insights

of these models, their institutional requirements are beyond what was available to societies in antiquity, and we will not develop them further here.¹⁰

In this section we will consider the contribution of skill development to economic growth. The first model uses the Solow model, with a Cobb–Douglas production function with capital (K) and skilled labor (H) as inputs: $Q = K^\alpha (AH)^{1-\alpha}$.¹¹ Individuals in the economy produce human capital by devoting time to learning skills instead of working. Letting L represent the total amount of raw (unskilled) labor in the economy and u be the fraction of a person's time spent learning new skills (a parameter rather than a choice variable), the production of human capital is $H = Le^{\psi u}$, where e is the base of the natural logarithm and ψ is the elasticity of H with respect to u . The capital accumulation equation remains the same, and the per capita production function is $q = k^\alpha (Ah)^{1-\alpha}$, where $h = e^{\psi u}$. Without repeating the solution of a model quite similar to the basic Solow model, the equilibrium output per worker is $q^*(t) = [s/(n + a + d)]^{\alpha/(1-\alpha)} h A(t)$. The key insight of this model lies in the influence of h on per capita production: a larger proportion of time spent in training will generate a higher per capita income.

A variation on this specification of skill accumulation offers a few more insights. Begin with a Cobb–Douglas production function quite similar to the one used in the previous human capital model: $Q = K^\alpha (hL)^{1-\alpha}$, where h is an individual's skill level, which enters this production function similarly to the Harrod-neutral technology in the basic Solow model. An individual uses time to create additional skills, but previous skills he possessed himself, as well as the generally available “frontier” technology also contribute to his accumulation: $\dot{h} = \mu e^{\psi u} A^\beta h^{1-\beta}$, where e , ψ , and u have the same interpretations as in the previous model. This activity does not produce the total quantity of skills the individual possesses, but rather adds to what he already has, and it is produced in a Cobb–Douglas form of production function in which he draws on the frontier level of technology, A , and his previous skills. This production function has the further

implication that the closer a person's skill level approaches the technological frontier, the more difficult it is to add more to his current skills. Rearranging the previous expression to take the form of a growth rate, we get $\dot{h}/h = \mu e^{\psi u}(A/h)^{\beta}$. As h approaches A , the growth rate of skills slows down, reaching $\mu e^{\psi u}$ when $h = A$. The solution for equilibrium output per capita becomes $q^*(t) = [s/(n + a + d)]^{\alpha/(1-\alpha)}[(\mu/a)e^{\psi u}]^{1/\beta}A(t)$, the first term of which is the basic Solow result. The second term corresponds roughly to the second term of the previous human-capital accumulation model, and indicates that countries in which people devote more time to skill accumulation will be wealthier. The last term represents the influence of the world technological frontier on any country; its effect on per capita production (income) will be mediated by the first two terms, which contain the effect of the saving/investment rate and growth rate in the country, and its acquired skill level. If we dig back inside the disaggregated version of the production function, we find the implication that different countries will use different types of capital, varying in sophistication appropriate to the skill levels of their labor forces.

Despite low growth rates in antiquity, and much of the technological advance probably having had the character of random, accidental discoveries, particularly in the earlier millennia, these versions of endogenous growth models point to explanations for different levels of wellbeing in different countries or regions of antiquity beyond accidents of natural resource endowments. Time devoted to developing skills would have depended at least partly on the institutions of the time and region, but where those institutions generated more highly skilled work forces, the capital equipment also would have tended to be more sophisticated – even though much of it may appear strikingly simple to us today. Nevertheless, much of the technological knowledge employed in any particular country at a given time would have been selected from an internationally available menu of technologies, a notion which archaeological evidence tends to support. People at any given time and place

picked out the equipment they could figure out how to use, and possibly to make. They very well may have been aware of other, more sophisticated equipment but lacked the skills to produce it, use it, or both.

Since models of endogenous growth involve systematic research and development, it is appropriate to note here the likely period of royal patronage of the Museum (Musaeum) of Alexandria in the development of water-lifting technologies, probably between 260 and 230 B.C.E., which resulted in the *saqqiya* and the *noria* (Wilson 2002, 8, 10, 32; endorsing Lewis 1997). Between the first and third centuries C.E., terracotta pots replaced the more expensive wooden buckets on the *saqqiya*, and Diocletianic tax incentives encouraged private investment in irrigation technologies, both fostering the adoption of these technologies (Wilson 2002, 9). The loss of some of the mass-production technologies in the Late Antique period and subsequently appears to have resulted from the contraction of demand, providing an alternative route to endogenous technical change: fired bricks north of the Alps and the potter's wheel in post-Roman Britain (Wilson 2002, 32). As to attitudes toward the application of scientific and technical knowledge in Greek and Roman antiquity, Cuomo (2007) discerns considerably more personal and social energy devoted to those activities than much of the surviving literary evidence would suggest. Her analysis of carpenters in Rome infers political tensions between trained, nonaristocratic groups and aristocratic, governing groups (98–102), a tension that still exists between technical experts and legislators. Focused as she is on teasing out the existence of these groups of possibly well-organized technical experts, Cuomo does not address their organization for passing on existing and expanding knowledge between generations. Nonetheless, her research shines light on the continuing interest in acquiring, using, and developing technical expertise in a variety of fields. Considering the growth in engineering knowledge in Pharaonic Egypt and the ancient Near East, it would not be surprising if evidence of similar government support (other than in trial-and-error

construction) appeared – unless I am simply unaware of it.

14.3.4 *Extent of the market, division of labor, and productivity*

Adam Smith's famous observation on the pin maker, loosely applied, suggests that productivity will increase with larger markets.¹² Thus, an ancient economy expanding its market through trade could be expected to sell (and produce) disproportionately more as the market for its products grew. What exactly does this mean? As an economy's demand curve shifts to the right as more customers want to consume its products, each productive unit, human and nonhuman, in the economy can produce more than it did when fewer people wanted its products.¹³

How can this happen? What are the mechanisms that produce the effect? Is it true? First, think about the basics of production theory, the production function. There is nothing in a production function that links the productivity of individual factors – equipment and labor – to the quantity of output; that would put the dependent variable, output, on both sides of the relationship, not an impossible theoretical structure, but a rather mechanical one once all the substitutions were made to put the dependent variable on one side of the equal sign and the independent ones on the other. Thus it is not particularly productive to begin thinking about the phenomenon at the level of the production function, although that will be a formulation to which to return.

The relationship between size of market and productivity is a market structure phenomenon – how firms are organized and how they compete with one another. When a firm produces goods that, say, are made of several components, but serves a small market, the firm will not find it profitable to specialize in production of one of those components and sell its output to an assembly firm that buys the components from several more specialized firms. The fixed costs would be too high. Thus we would find vertically integrated firms. As the firm's market grows (and we can think of this both over time and as atemporal snapshots of different production levels), it will

hire more employees and acquire more pieces of equipment (think of short-run average cost curves along a long-run average cost curve). At some point in its scale of production, it can assign workers to more specialized tasks and acquire machines that they can work with on those tasks. The increasing productivity per unit of factor of production contributes to the downward slope of both short- and long-run average cost curves. However, as output continues to increase, the costs of coordinating those more specialized factors of production increase, eventually making it profitable for the firm to split itself into several independent production units, each a separate firm. Each of those firms will employ workers who produce a smaller array of products, thus allowing them to become more skilled (efficient) at their tasks, and similarly to use more specialized pieces of equipment. Since there are fixed costs to both inventing new machines and to labor's acquisition of higher levels of skills, larger scales of production facilitate both activities. One prediction of the theory is that vertical integration of firms will decrease as the overall market size increases.

Of course, part of the productivity effect comes from technological change: the more specialized equipment used by the more specialized firms. Whether this effect represents genuine technological change in the sense of new knowledge being built into new equipment, or just existing knowledge being applied to slightly different capital goods that find narrower application, will govern the magnitude of the effect of vertical dis-integration accompanying the expansion of a market.

Returning, as promised, to the level of the production function – rather than simply specify factor productivity as a function of output, which in fact was the phenomenon to be explained in the first place, analysis of specialization has focused on the array of tasks within a production process and the distribution of skills among individual units of factors (say, labor) that lead to comparative advantages of different units of factors in conducting specific tasks. At this level of analysis, the specialization that emerges depends on relative prices of skills and relative

values of tasks but ultimately does not depend on scale of production. What does depend on scale of production is the distribution of skills among the labor force, a greater demand for the products of particular skills reducing the risks of investing in those skills. And similarly for specialization of equipment.

It's a nice theory but does it hold any water? By and large, yes, but in applications to antiquity it may be useful to consider the scope for the impact of this phenomenon on long-term growth of product per capita in an economy. If, to heighten the example, the vertical dis-integration that permits greater specialization and hence higher productivity were to occur all at once, we would see a lengthy period of flat (that is, no) productivity growth, a discrete jump in productivity at the time the firms split apart, and a subsequent, higher level of productivity but still with flat productivity growth. More realistically, such changes would occur gradually over time, and without increases in the knowledge base underlying both people (human capital development – through general education and specific training, both of which are costly) and equipment, growth would eventually level off.

14.4 Structural Change

The one-sector growth model does not let us study how the composition of activity in a society changes as growth occurs. The study of economic development, conceived as a process of individual and social learning, increasing sophistication, and division of labor, needs some other models, which fortunately exist in plenty. There is no single model for all occasions, but rather a set of principles to be used in constructing any particular model to study a specific case. First we introduce the concept of the sector, which is simply an organizing and accounting unit in these models – a method of organizing our thinking about a particular problem, according to certain principles, for purposes of analysis, rather than an organization that the agents at the time and place under study might have recognized themselves. Next we sketch the outlines of a simple,

two-sector model to show what goes into it and how the various parts relate to one another. The section closes with an overview of some empirical regularities at the level of the entire economy from the nineteenth and twentieth centuries, some of which surely apply to entire economies in antiquity, others of which equally certainly differ, but in predictable ways.

14.4.1 *Sectoral concepts as organizing devices*

A sector is an accounting convention that assigns similar activities, usually production, to a common accounting group. Purposes of a particular investigation may make it useful to throw together different activities that share some common characteristics and separate them from other activities that tend to have different characteristics.

It is common to see distaste expressed in historical and archaeological writings for thinking of economic activity in terms of *sectors*. After all, most of the population farmed for most of their livelihood, but many, if not most, of the industrial and craft activities were conducted by the same people in their agricultural “down-time.” The same has been true in recent times in the countries called the Third World today, and even in the Western, industrialized countries into the nineteenth and even the early twentieth centuries. Fortunately, the concept of the sector of economic activity does not rely on separate people doing different things. All the sectoral concept does is recognize that different activities have different production functions and that people who produce in one sector generally consume the products of other sectors as well, which ties the different activities together so that neither is truly independent of the other.

Sectoral models are useful when we want to study situations in which the existence of and behavior in one major part of an economy constrains or otherwise affects events in another major part. Common subjects of study are urbanization, industrialization, economic growth, and economic development. The concept of sector is useful for identifying major groupings of

economic activity that use substantially different technologies, have different consumption characteristics, or both. Both technological and consumption differences will yield definite implications for transfers of economic resources between them.

For example, consider an economy using two factors of production in two sectors of activity. One of the sectors employs 10 units of labor for each unit of capital it employs, while the other uses 5 units of labor to each unit of capital. Suppose some outside event occurs – say, trade begins with a new country, increasing the demand for the output of the first sector, signaled by an increase in the price of sector 1's output relative to that of sector 2. The labor and capital are fully employed between the two sectors before the demand increase occurs. To produce more of sector 1's output, that sector has to find labor and capital, and the only place to get any more is from sector 2. Through the inducement of higher remuneration, sector 1 receives some capital and labor from sector 2, but for every unit of capital it gets, it gets only half the number of workers required to operate sector 1's production process. Stated alternatively, for every unit of labor it gets, it has an extra unit of capital. From sector 2's perspective, if sector 1 takes labor and capital away from it at the rate of 10 to 1, sector 2 will be left with more capital than its workers can operate. To maintain full employment (and unemployment, defined as doing absolutely nothing, on a permanent basis, is simply not an option, either theoretically in this type of model or in the real world of either modernity or antiquity), the input ratios in both sectors' production processes have to change. To accommodate those changes, the ratio of labor payments to capital payments must change. And we must not forget that the revenue from exports of product 1 must be spent on a corresponding value of imports to maintain external balance, reinforcing the reduction in sector 2's output. The expansion of sector 1 is constrained by its ability to take resources from sector 2, and that in turn is constrained by the country's continuing, if reduced, demand for sector 2's product.

Two common sectoral distinctions are the agricultural / industrial and the urban / rural dichotomies. There is considerable overlap between the two, but they emphasize different topics. Agriculture is associated with "rural," but some urban dwellers farmed in antiquity (and into recent times as well, of course). Industry can be associated with cities, but much industrial activity occurs in the countryside. "Urban" brings the connotation of industry but also of services – doing things for people rather than making things they can eat or wear. Also population concentration and agglomeration and scale economies of various types typify urbanism. Similarly for consumption: housing and land take on a prominence in cities that they do not have generally in rural areas where space is less scarce.¹⁴ Quite important among consumption differences between urban and rural sectors is the greater frequency of interpersonal interactions facilitated by higher population concentrations. This type of consumption brings with it various sorts of learning which, altogether, generate greater sophistication and demand for different arrays of consumption goods.

Different factors, or factor ratios, are associated with the different sectors, both in the technological requirements of the production processes and in the quantities available. The industrial-agricultural dichotomy generally emphasizes higher ratios of capital to labor in the former sector. Sometimes different sets of factors are used for the two sectors: capital and labor in industry and land and labor in agriculture. Such a characterization does not, of course, imply a belief on the part of the student that industrial activity is conducted on absolutely no land, but that land is such a small proportion of input costs that it can be ignored safely for many purposes. Similarly with the exclusion of capital from agriculture. If we happened to be interested in studying what the effects could have been of a growing capital stock in agriculture – including animals as well as implements, we certainly would include capital as an input in agriculture.

Correspondingly, in the urban-rural sectoral distinctions, it is difficult to ignore land in cities because so much of the distinction in activity

structure in cities is attributable to the competition for scarce land. Thus the typical urban sector specified in a model would use labor, land, and capital, while the rural sector might be restricted to just labor and land, with the specification of capital being a luxury to the modeler. Again, the questions of interest and the solution tools available influence these modeling choices. Of course, when we move from the results of a model back to evidence from the observational world, of either modernity or antiquity, we need to remember the omissions we've made in the modeling. Thus, although we might have omitted capital from agriculture for good reason in our modeling, when comparing model results with empirical observations, we need to be cognizant of the possible influences the omission of capital from agriculture might have on departures of predictions from observations.

14.4.2 *A two-sector model of an economy*

We could model an economy as having any number of sectors we wanted – a three-sector model, such as industry, services, and agriculture; or a multisector model that distinguished between different agricultural activities as well as different industrial and craft activities – but the two-sector model demonstrates all the interactions and their importance and is simple enough to let us see how much of the interaction works. We will use the agricultural-industrial distinction.

Five components are required to assemble any general equilibrium model: production, the demand for factors of production, the demand for outputs, the employment of factors, and market balance which equates total production to total demand. The production component consists of two production functions, one for the agricultural good (a composite good, of course) and one for the industrial good (another composite). Let's suppose that agriculture uses labor, capital and land; then its production function would be $Q_A = f(N_A, K_A, L_A)$. Suppose that industry uses only labor and capital: $Q_I = g(N_I, K_I)$. We use f and g for the production functions to indicate that different technologies are used

in the two sectors. That's all there is to the production component.

Remember that the demand for a factor of production is determined by the value of its marginal product. Then the wage payment per unit of labor in agriculture is $w = pMP_{NA}$, where p is the relative price of the agricultural good in terms of the industrial good, or p_A/p_I , and MP_{NA} represents marginal product of labor (N) in agriculture (A). The wage in industry is $w = MP_{NI}$. Notice that we have let w represent the wage in both agriculture and industry, the implication of which is that enough labor is able to move between the two sectors to equalize the wage. The demand for capital in agriculture is $r = pMP_{KA}$ and the corresponding demand in industry is $r = MP_{KI}$. Again, letting r represent the rental rate in both sectors indicates that capital also can be moved between sectors to equalize the return to capital. Finally, only agriculture has a demand for land: $z = pMP_{LA}$. This notation may give the impression that one of these marginal products could change independently of the others, but remember that the marginal product contains within it the average product, which is output divided by labor, capital, or land. If anything alters the marginal product of one factor in one of these activities, the marginal products of all the factors in both their employments will have to adjust: these factor demand equations are all tied together.

There are several ways to define the demands for the sectoral outputs, depending on how we allocate income. We have income deriving from labor, capital and land, and we must characterize its allocation across groups of individuals. How we do so depends on what we think about institutions. We could distinguish radically between the people who own capital and land on the one hand and workers who own only their own labor (if that). Alternatively, we could allocate the income from capital, land, or both equally among all workers. Any combination of these allocations is possible, and through specification of how much of each source of income each group of income recipients saves, we could set the stage for a changing time path of income and wealth among social groups. But to return to the immediate task

of specifying some general forms for commodity demand, let's suppose that workers own only their own labor (but they do own all of that) and that another social class owns all the capital and all the land. We need workers' demands for the agricultural good and the industrial good: $D_{NA} = N \times d_{NA}[p, (1 - s_N)w]$, which says that N workers have the same demand function for agricultural goods, represented by d_{NA} , which is a function of prices and the income they choose not to save (s_N is the workers' saving ratio, which could be zero). We must specify a corresponding workers' demand for the industrial good, D_{NI} , which would have the same general form but a different functional relationship designated by d_{NI} .¹⁵ Since we haven't specified the actual people in the class of capital and land owners, we will designate their demands as aggregates. Their demand for the agricultural good will be $D_{CA} = d_{CA}[p, (1 - s_C)\gamma]$, where $\gamma = rK + zL$ is the total income from capital and land.

We are not quite finished with the demand component, because we haven't disposed of the income that people allocated to savings. The supply of savings equals the demand for investment. Investment is what replaces worn-out capital and adds to previous stocks if accumulated in sufficient quantity, so it is reasonable to suppose that investment is a particular nonconsumption disposition of the industrial good. The demand for the industrial product as an investment good is $I = (s_C\gamma + s_NwN)/p$, in which dividing by p gives the property that the demand for the investment good falls as its price rises.

The next component is the demand for factors, which ensures that full employment of both labor and capital is maintained. We don't need to worry about land since it has only one use, and any return it earns is a pure rent. For labor we have $N = N_A + N_I$, which says that the total amount of labor in the economy, N is allocated fully between agriculture (N_A) and industry (N_I). For capital we have the equivalent adding-up: $K = K_A + K_I$. Keeping track of these two conditions keeps us from either thinking we can use more of one of these resources than the economy has or forgetting to use all there is.¹⁶

The product market balance forms the next model component: supplies must equal demands, in physical units. For the agricultural good, $Q_A = D_{NA} + D_{CA}$, and for the industrial good, $Q_I = D_{NI} + D_{CI} + I$.

A sixth component is required to govern the behavior of this economy over time. This component contains specifications of how technical change behaves, how or whether investment turns into capital accumulation, the growth of population, and how the land supply may be augmented. We have not written the state of technology into the very general production functions above, but the previous section discussed several methods of modeling technical change. Accordingly, we could just specify the rate at which we believe technological progress occurs(ed), and of what sort. For capital accumulation, we have the relationship from the previous section that the absolute growth of the capital stock is investment minus depreciation: $\Delta K = I - \delta K$, where δ is the depreciation rate. The growth of the labor supply is simply the annual growth rate times the current year's population: $\Delta N = nN$. We could specify the growth of arable land as either an exogenous increment or make land clearance another activity requiring labor and capital.

What sort of insights can models of this sort yield us? One result has important implications for urbanization. As long as technology permits each agricultural worker to produce sufficient food for, say, only 1.11 people, 90% of the work force will have to remain in agriculture. If cities attract predominantly nonagricultural workers, the urbanization of the entire economy is governed by the same ratio. If some city residents farm, loosening the overall percentage of the population urbanized, their numbers and agricultural output still will be constrained by access to land, and the economy's urbanization is still fairly tightly constrained. Only technological change in agriculture will loosen this limit on urbanization.

14.4.3 Some stylized facts

The record of the currently industrialized countries over the past several centuries, as well as that

Table 14.1 Composition of labor force and aggregate income, and urbanization.

Country	Year	Agriculture ^a	Manufacturing	Services	Percentage urban
<i>Percentage of labor force</i>					
Taiwan	1905	72.0	5.0	21.9	3.50
Brazil	1920	67.0	13.0	20.0	3.23
Chile	1907	37.7	17.6	41.9	24.2 ^b
Japan	1872	85.0	6.0	9.0	4.96
Japan	1925	52.0	24.0	24.0	11.18
United States	1810	80.9	2.8	16.3	6.06
United States	1840	63.1	13.9 ^c	23.0	10.81
<i>Percentage of gross domestic product</i>					
Korea	1910–12	91.8	6.7	–	3.3
India	1900	57.7	11.9	30.4	2.03

^aIncludes fishing and forestry.^b1912.^cIncludes construction of 5.1%.

Sources: **Taiwan:** employment shares, Ho (1978, 82, Table 5.4; 326, Table A23; 332, Table A25); urbanization, Mitchell (1982, 56, Table B2; 67–71, Table B4). **Brazil:** employment shares, Katzman (1977, 152, Table 21); urbanization: Mitchell (1983, 51, Table B1; 105–110, Table B4.) **Chile:** employment shares, Mamalakis (1975: 11, Table 1.2); urbanization, Wilkie and Perkal (1984, 140, Table 653). **Japan:** Kuznets (1966, 107–108, Table 3.2); urbanization, Mitchell (1982, 56, Table B2; 67–71, Table B4). **United States:** labor force, Lebergott (1976, 119, Table 2); urbanization, Bureau of the Census, U.S. Department of Commerce (1975, Series A34–50). **Korea:** labor force, Kuznets (1977, 19, Table 1.4); urbanization, Kim and Sloboda (1981, Table 13). **India:** production, Maddison (1971, 167–168, Table B-1; 169, Table B-2); urbanization, Mitchell (1982, 56, Table B2; 67–71, Table B4).

of the Third World countries over much of the past century, shows some structural regularities among major categories of economic activity and between the shares of economic activity in those categories and urbanization. Table 14.1 shows these trends for a variety of countries. As the share of either the labor force or aggregate income deriving from agriculture falls – either across countries or within a single country over time – the share from manufacturing naturally rises, but more important to notice is that the share in services also increases. Frequently the increase in the share of income coming from, or labor force in, services is greater than the increase in manufacturing. The decrease in agriculture's share is attributable to some combination of technical changes in a country and increased imports (which implies a more productive agricultural technology in the exporting country); otherwise the people shifting their efforts to manufacturing and services would not be fed. Of course, some of the manufacturing and service activities are

conducted in the countryside but an increase in urbanization accompanies this shift away from agriculture. An increase in urbanization implies a decrease in agriculture both in terms of people's participation in it and as a share of total output, not simply a decrease in rural living. Similarly, an increase in manufacturing implies an increase in urbanization, not simply a change of activities in the same villages. Also, as people shift from agriculture to manufacturing and from countryside to city, they consume more services provided by others—an increase in the division of labor.

Table 14.2 averages sectoral shares from a number of developing countries and compares the income differences associated with the differences in structural patterns, information missing from Table 14.1. The activity sectors are divided into primary, which includes fishing, forestry, and mining as well as agriculture; industry, which includes construction in addition to manufacturing; and services, which includes transportation

Table 14.2 Changes in structure of production as economies develop.

<i>Sector</i>	<i>At lower income level</i>	<i>At higher income level (4.3×)</i>
<i>Percentage of gross domestic product</i>		
Primary	52.2	26.6
Industry	12.5	25.1
Services	30.5	48.2
<i>Percentage of labor force</i>		
Primary	71.2	48.9
Industry	7.8	20.6
Services	21.0	30.4
<i>Percentage of population urban</i>		
	12.8	43.9

Source: Chenery (1979, 12–13, Table 1.1).

as well as a range of more personal services. The middle column of the table shows the percentages of national income and labor force in each sector at the lower level of income, and the rightmost column shows those percentages at an average income level 4.3 times higher than the average income of the countries in the poorer group. Clearly, few if any ancient economies ever saw their share of agriculture in either output or labor force drop as low as 52%, although some small countries known as traders, such as Middle and Late Bronze Age Ugarit and the later Phoenician city states, may have approached such a structure. One interesting pattern in this table is that the share of the labor force in agriculture and related activities is considerably higher than the share of income deriving from that sector. Workers are more productive in industry and services, the predominantly urban activities. It is easy to see that at both income levels, the proportion of agriculturalists living in cities need not be high: the percentages of the population that are urban are smaller than the percentages of the labor force in industry and services.

Undoubtedly the percentages represented in these activity and urbanization patterns differ between the countries and times represented in these tables and the countries of the Mediterranean and the Aegean in antiquity, but the

directions of association contain some imperative elements. People cannot leave agriculture without those remaining behind becoming more productive. While there is some joint causation here, the capability to become more productive is the dominating force. The increase in urbanization as manufactures gain in contribution may have been smaller in antiquity than in recent centuries but it is difficult to interpret the ancient cityscapes of, say, lower Mesopotamia in the late third and early second millennia as housing anything close to the majority of those lands' agriculturalists. It is doubtful if those Sumerian and Early Babylonian cities of whose remains we are aware, if not all of them have been explored (and few in much detail), could have contained more than about 10 or 15% of those countries' populations. The agricultural and transportation technologies simply would not have permitted it. Similarly, while it is easier to see the remains of some manufacturing activities, especially metals manufacturing, services – inns, taverns, porters, doctors, lawyers, food preparers (to avoid the term restaurants) – are more difficult to see in the artifactual evidence, but they must have accompanied the growth of manufacturing, especially in cities.

14.5 Institutions

Institutions affect the performance of an economy by their effects on the costs of production and exchange. A major economic role of institutions is to reduce uncertainty by establishing a stable, but not necessarily efficient, structure to channel human interaction. Their stability does not obviate the fact that they change, albeit commonly in slow changes, "at the margins." Institutions are created by people, and they are altered by people as the source of their development (North 1990, 5–6). One common criticism of economics is that it ignores institutions. While there is much substance to that charge, the past quarter-century has seen a good deal of research on the economics of institutions. Tracing back from what institutions accomplish and who benefits, research has focused on the motivations to form

institutions, to maintain them, and to alter how they operate.

Douglass North, an economic historian who shared the 1993 Nobel Prize in economics with Robert Fogel, characterizes three building blocks of an economic theory of institutions as a theory of property rights that describes individual and group incentives in the system, a theory of the state, since the state specifies and enforces property rights, and a theory of ideology that explains how different perceptions of reality affect the reaction of individuals to changing “objective” situations. His ideal theory of the state would explain the inherent tendencies of political-economic units to produce inefficient property rights and the instability of the state in history (North 1990, 7–8, 17). These are tall orders, and it is fair to characterize the literature on this topic as an active research area.

14.5.1 Property rights

Property rights give the right to exclude others from access to or use of resources, and they may be conferred by law or custom, frequently with interaction between the two sources. Property rights reduce uncertainty and make a wide range of efforts more worthwhile. Rights to physical property, from land and housing to farm tools and food, are necessary to the investments required to produce these items in quantities that confer a steady and high stream of benefits.¹⁷ More complicated physical assets require more intricate property rights: more people are likely to be involved in their production, finance, and use, and there are more aspects of them to be used. Financial assets, either implicit or explicit, typically accompany larger physical assets, and they are a means of implementing property rights. Defining property rights requires identification and measurement of what is to be defined, as well as a technology adequate to store documents identifying rights-holders. And to be effective, the cost of enforcing property rights must be lower than the benefits of retaining them, although some losses to various sources of infringement may still occur.

Of equal importance to economic growth are rights to intellectual property, the ideas underpinning the technological change required for sustained growth. Until recently, the people who produced profitable ideas were unable to benefit systematically from them; they had to rest content with recognition and sporadic awards or prizes. While such rewards are not to be discounted, they are not capable of giving the incentive to engage in sustained investigations and dissemination of findings. Of course, the development of science as the experimental and public activity that it has become in recent centuries was a complex process, and I do not mean to imply a simple grafting of the patent concept onto the processes and events that involved religion, politics, and philosophy in the creation of modern science. Nonetheless, one potent factor holding back technological change in antiquity, and possibly even blunting the incentives for the sorts of investigation we call science, was the absence of the ability of an idea’s originator to command payments from others using it. Palatial control of technological information, established as well as new, could have been an effort to obtain some of the benefits conferred by patenting but without the information-management infrastructure available in recent centuries.

14.5.2 Governments

A useful definition of a government, or a state, is an organization with the authority and capability to collect taxes over some geographic territory. Possession of a comparative advantage in violence, or force, confers on it the capacity to enforce various laws, including property rights. A government is also the vehicle people choose for specifying property rights as well as supporting enforcement of them.

Recent economic theories of government emphasize the provision of services, including protection, in return for revenue. The monopoly, or at least the comparative advantage, in violence leaves the state the agency capable of offering protection to constituents, as well as compelling their support, if not necessarily their unfailingly enthusiastic participation. It is easy to overlook

this exchange element in ancient governments when the records of the ancient governments themselves frequently emphasize the religious character of kings. Notice or recall, however, that religion did not protect kings who failed to deliver what subjects, either aristocratic or common – peasants – expected. Regicides of and coups against kings who were believed to embody various divinities occurred from time to time and were accepted by their constituents. Even the Intermediate Periods in Egypt can be interpreted as times when the exchanges involved in governing were executed with less competence than usual, or were interfered with by outside events with the same result as incompetence.

Since the state has a particular interest in military technology, weaponry commonly is at the forefront of technical change. The lead role of warships in the development of marine technology is a good case in point. It is likely that secrecy, rather than any form of open disclosure, would have characterized systematic development activity in this field, which need not run counter to the evidence of ancient, international weapons sales.

The resource allocation choices that governments make, from the particular property rights they are willing to specify and enforce, to whom they tax and buy from, are the outcomes of competitions among various interest groups. Sheer size of populations in interest groups frequently works against their effectiveness in lobbying, since the benefits that large groups persuade the government to transfer to them from others would have been shared too widely to be very sizeable for any individual. These choices of the state affect, often retard in fact, efficiency, but they equally often are not designed to improve efficiency in the first place.

14.5.3 *Stability and change*

The state provides its constituency with a certain measure of stability, even if at the cost of some repressiveness. Stability reduces uncertainty and permits planning over longer time horizons and in greater detail. Economic agents can adjust to a predictable “bad deal,” but not knowing whether

the state will permit a “good deal” or any of a number of “bad deals” to occur would hamper planning and possibly drain resources into more planning than would be necessary under a known “bad deal.” More capital-intensive projects can be undertaken under greater stability, which translates into more investment and consequently into higher consumption per capita. Of course, stability does not imply efficiency.

Actual institutions of a state may outlive the current ruling family or party. They can survive coups and revolutions, leaving an underlying degree of stability facing the majority of economic agents even if the superstructure of the state displays a marked degree of instability. Of course, if policies of successive rulers differ greatly from one another, the economy is bound to be affected by reversals and attendant periods of unpredictability.

Institutions and governments are always changing, but generally at “the margins” and in response to changing incentives of individuals either inside or outside the government. They can also change, sometimes permanently, in response to temporary external crises such as climatic events or invasions, but typically the majority of government functions will remain much the same over a period of a human generation.

Governments also are instrumental in providing instability as well! While stabilizing policies have fostered periods of sustained economic growth in a number of countries in the past two centuries, and surely the same can be said for a number of lands and periods in antiquity, the pursuit of bad policies can put an end to prosperity, not to mention growth. Confiscatory and arbitrary tax policy, wars, and inflationary monetary policy in monetized economies are the usual lineup of culprits for throwing economies into disarray from which they often have difficulty recovering.

14.6 **Studying Economic Growth in Antiquity**

Much of this chapter has been pretty abstract. This section uses as a backdrop current interest

in the subject of economic growth in antiquity to suggest some ways to structure analysis of the subject. The first thing to think about is how much there could be to explain. The second subsection suggests an organization for thinking about the topic in terms of outputs and inputs, which I think may help students of the subject distinguish consequences of growth from causes, which are not that difficult to conflate if one is not careful. Since the level of wellbeing at the end of antiquity probably wasn't a lot higher than it was at whatever one decides to consider as the beginning, the next subsection offers suggestions for studying fluctuations in growth, not as business cycles, which are a subject beyond what can be treated presently, but of the longer-term responses to external shocks. A short summary section pulls together the principal ideas of the section.

14.6.1 *What there is to explain*

Interest in economic growth in antiquity has surged since the turn of the millennium. One fact facing this research is that the economic growth of the northwestern European Industrial Revolution is known to have begun sometime in the seventeenth century C.E. from a fairly well-understood base of consumption or output per capita. The tacitly accepted retrenchment in output per capita beginning in Late Antiquity and extending through the European Middle Ages provides the well-accepted trough in prosperity between a prior peak sometime around the second century B.C.E. to the second century C.E. (although the dates of any such peaks may vary among locations). Discoveries of evidence for economic growth during antiquity face an empirical ceiling on the prosperity that growth could have produced, of something not too much higher than, if not quite as high as, the standard of living at the beginning of the Industrial Revolution. So despite all the apparent evidence for growth at various places and at different times, any growth that occurred couldn't have been too extensive.

Working backwards in time and focusing on Greece, it is widely accepted, if subject to changing nuances, that output per capita reached a nadir

sometime in the Greek Dark Age that followed the collapse and / or decline of the Mycenaean civilization – or Kingdom of Ahhiyawa (Kelder 2010). Working backwards still further, the Mycenaean states appear to have reached pinnacles of prosperity beyond those achieved during the Middle Helladic Period, which in turn appears to have been less prosperous than a number of places and times during the Early Helladic (Voutsaki 2010). So, there are a number of possible periods for examination, but the generally accepted pattern of broad temporal waves of prosperity and retrogression over some three millennia indicates the clear scope for distinct periods of nonzero economic growth. The period from around the eighth century B.C.E. through the second or third century C.E. has attracted the most attention recently, possibly because it provides the most evidence.

14.6.2 *Organizing inquiry about economic growth with the help of growth theory*

With these numerous episodes in mind, and bearing in mind also that other regions of Mediterranean antiquity will have had comparably interspersed episodes of prosperity and recession or depression,¹⁸ it is not true that every increase in output per worker or improvement in standard of living is necessarily attributable to the influences modeled in economic growth theory, which uses a production-function approach to relate output to increases in the quantities and quality of inputs. As discussed below, changes in relative prices and increases in technical efficiency also can increase output, although the increases they provide will be one-time changes, even if they occur over time. Recovering technological knowledge lost during a previous period of disorder provides the mechanism modeled in growth theory, but an increase in prosperity associated with such a rediscovery will only return living standards asymptotically to the previous level without continued expansion of the technological frontier through new discoveries.

Nevertheless, some concepts from growth theory can help organize questions regarding

manifestations and causes of observed or putative growth. Start with the fruits of growth, because that's where the bulk of the evidence will probably lie. Define growth, from the perspective of growth theory, as a sustained increase in output per worker, which translates into a sustained increase in utility per worker.¹⁹ Actually, a prospective increase in utility is the incentive for the actions that produce the increase in output per worker. The other end of the problem, where the growth comes from, involves technology: capital per worker and the technology embodied in the capital.

The output side

On the fruits-of-growth side, growth theory typically deals with a single, undifferentiated product or at most a few products, whereas many products exist in the observational world. Production of some goods may increase on a per capita basis while production of others might remain constant or contract. The variety of products available may increase even if the output of pre-existing goods remained constant or decreased. Locational differences in tastes and input costs can contribute to these variations in experience. We can try to look at what is produced or what is consumed, but it is useful to keep in mind the difference. Increases in food production per capita certainly would have been a major component of economic growth during this period, but some of the most important evidence for increases in agricultural productivity will come from skeletal evidence of improved nutrition, which is on the consumption side of the equation, and other forces can affect nutrition: health care (which probably changed little over the period), urbanization, and changes within agricultural production itself, to identify three. One recent observation to keep in mind, however, is John Komlos' finding that the stature of twentieth-century C.E. Americans has fallen behind that of northwest Europeans, a striking observation which has no satisfactory explanation currently (Komlos and Baur 2004). Pilkington (2013, 7–8) reminds readers that reference to adult skeletal remains without consideration of evidence of possible

nutritionally induced stunting and ill health among infants and juveniles can be misleading: a population with high infant and child mortality rates still can produce tall children who are most likely to reach adulthood. Consideration for students of antiquity: be careful when attributing changes in stature to corresponding changes in material wellbeing. Both Komlos and Baur's and Pilkington's observations highlight the potential pitfalls of interpopulation comparisons of heights, among either the living or osteological remains.

Observations of consumption may mix together changes in the distribution of income as well as possible overall increases in income. Wealthier people may be more visible archaeologically than the humbler members of a society, and the appearance of more wealth among the observable remains may give the appearance of considerable growth that is actually partly or wholly redistribution of income from poorer workers who produced the goods to an aristocracy which was able to enforce claims to large parts of the production. The standard of living at the level of the typical worker puts a limit to the extent of income redistribution, as opposed to economic growth, that could have been involved in such observations. Although the concept of a minimum living standard is elusive – people can fail to die, and even continue to reproduce, on widely varying nutritional intakes – there surely would have been a limit to which an aristocracy could have shifted enforceable legal title to a largely unchanged volume of production without killing the geese that laid the eggs.

Architectural remains of private houses may give some insight into changing wealth, but care must be taken to compare similar surroundings. Buildings well inside a heavily built-up urban area tend to be more compact than buildings with comparable purpose located in rural areas or less heavily built-up areas on the edge of a city. Land prices are influenced by city size and are a constraint on the horizontal expanse of dwellings. Occasionally evidence of upper stories is available in architectural remains, which could indicate square footage in excess of a building's footprint. Furthermore, if architectural remains

from lower income strata of a society are less readily visible archaeologically than are those from upper income levels, the data may be picking up an artifact of income distribution rather than of economy-wide economic growth. The ideal comparison from which to infer increasing wealth (and possibly economic growth) would be changing dwelling size in comparable areas of a single city. Such clarity in physical remains might be a luxury, leaving care to be taken in inferring changes in wealth from dwelling sizes at different locations and dates.

Great expansions of international trade have been pointed to as evidence of economic growth. Without changes in technology, changes in conditions that lead to increased trade take advantage of differing technologies and factor proportions among trading partners. People at both ends of the trading benefit with increased incomes. Once these advantages of trade have been fully exercised, no further improvement in income will occur without technological improvement, which could come in the form of the production technologies of the goods traded or in the transportation. Technological change cannot be inferred, and growth theory has nothing to offer to the explanation of increased wellbeing deriving from expanded trade.

The input side

Increases in capital per worker increase output per worker, but with diminishing increments of output as more capital is employed. Eventually, output per worker is maximized at an optimal capital-labor ratio. Adding more capital beyond that point would reduce output and consumption. Without any knowledge of the theory, people probably could figure out when they would have been using more capital than was productive. Empirically, physical capital accumulation has been found to account for a surprisingly small proportion of observed growth in output, estimates ranging from a fifth to a third.²⁰ This leaves technical change, embodied in the physical capital, or in the human capital in the form of new knowledge or new ways of combining inputs, as the principal long-term determinant of continuing economic growth.

People do ask, of course, why technical changes occurred when they did rather than earlier, and explanations tend to run in terms of attitudes and institutions, the latter a sufficiently elastic concept to include income distribution as it is affected by government policies. Institutional developments, such as the codification and enforcement of commercial law, typically improve the efficiency of resource allocation, but without continued technological change, those increases in allocative efficiency would provide one-time improvements in living standards.²¹ Sometimes, of course, even frequently, knowledge is cumulative, such that one invention cannot precede another, and then there are the typical lags between invention and innovation, which are fairly well understood for contemporary industrialized economies but are poorly observable, much less understood, for antiquity. Probability-related explanations for inventions rely to a considerable extent on the size of populations thinking about problems, with larger populations more likely to hit on solutions than smaller ones.

The size of populations, or at least of markets – allowing for the linking up of markets of separate populations of unchanging size – offers two more routes to raising output: scale economies and specialization. Scale economies of most ancient production technologies may have been quite limited but a larger population could permit larger scale public infrastructure creation, which enhanced productivity of private inputs. Irrigation canals and other large-scale waterworks and land transportation systems are prominent examples. Larger populations or larger markets permit individuals and firms to concentrate on a smaller array of tasks in production and may foster the development of new technology to work with these more specialized tasks, resulting in greater productivity – that is, more output from the same quantity of inputs.

The models dealing with these issues fall outside the basic, production-function-based models of growth theory per se, although sometimes they have been coupled with growth models to develop more comprehensive explanations of episodes of growth or nongrowth.

14.6.3 *Studying episodes of growth following declines: beyond growth theory*

Returning to the theory, it bears re-emphasizing that economic growth theory deals with growth in per capita, or per worker, output, not growth in total output. Once an economy has accumulated the optimal amount of capital per worker, with no change in technology, economic growth will be zero, although output and consumption per worker are at the maximum levels sustainable. An increase in population that was not matched by an increase in capital that kept the capital-labor ratio constant would generate negative economic growth – a decrease in output per worker.

It is important to distinguish between the consequences of events that change relative input prices on the one hand and technical progress on the other. A plague that substantially reduced a region's population without materially affecting its capital stock would increase the marginal product of labor simply because the capital-labor ratio increased, with no change in technology. People surviving the plague would indeed be better off than they were before, partly because the reduction in total output is less than the reduction in labor inputs since capital is unaffected, and partly because of a redistribution of income between owners of capital and suppliers of labor. If ownership of capital were distributed evenly among surviving workers, postplague real output per worker would be greater than under preplague circumstances. One way of seeing this is to suppose a Cobb–Douglas aggregate production function for a first approximation and note that the percentage reduction in total output from a percentage reduction in labor inputs would be the percentage reduction in the labor force multiplied by labor's output elasticity, or labor's share of the value of output, a number less than one; this yields an output reduction smaller than the labor input reduction. Next, for all practical purposes, the surviving workers can claim the capital left by those succumbing to the plague and appropriate the product of that capital. Furthermore, the increased capital-labor ratio would increase the marginal product of labor and

reduce the marginal product of capital. Relying again on a Cobb–Douglas approximation for aggregate production, the percentage increase in the marginal product of labor would be equal to the percentage reduction in the labor force times the negative of capital's output elasticity, or capital's share of the product, a positive result as the reduction in the labor force with a constant capital stock increases the wage rate relative to the rental on capital.

Additionally, given the unchanged technology, the postplague capital-labor ratio would be higher than the optimal capital-labor ratio consistent with preplague factor supplies. With a continuation of unchanged technology, saving would fall to reduce the capital-labor ratio to its optimal level: as capital wore out, investment would not fully replace depreciation. If the labor force remained at its postplague level, the capital stock would have to fall further than if the population grew somewhat. The temporary boost in the marginal product of labor occasioned by the plague's depredations would erode gradually until the preplague marginal product was reached.

How would the population respond to a sudden reduction in its magnitude? To some extent, this is a problem in demography but in the past several decades strides have been made in endogenizing population growth within economic growth models. Some of the efforts have remained at the level of interactions with aggregate population, others have incorporated utility-maximizing fertility choice models, and yet others, while stopping short of fully integrating demographics with growth models have modeled demographic-economic interactions in some detail.²² It is difficult to get all the effects one would like to know about in a single model and have it remain tractable.

Of the population models that have been studied in these interacting contexts, the supply-side Malthusian model is pretty simple, if fairly robust, but it does not rest on microeconomic foundations of fertility choice.²³ The principal alternatives are the demand side Becker–Lewis model of the demand for children (Becker and Lewis 1973), which relies on price and income, and the Easterlin model of fertility

(Easterlin 1966a, 1966b, 1978),²⁴ which appeals to an endogenous change in preferences (or expectations, depending on one's viewpoint) as birth cohort sizes change, within an overall economic approach. None of these models is necessarily wrong, although one may better characterize a particular situation than the other does, and in fact, the Malthusian model appears to perform quite well prior to the Industrial Revolution. In fact, the economic structure of the Malthusian model corresponds well to an income constraint as the major influence on fertility, with very weak relative price responses and no quantity-quality tradeoff in child choices because of largely nonexistent opportunities for human-capital expenditures on children.

Barro and Becker (1989, 491–492) incorporated an extension of the Becker–Lewis model of the demand for children into a growth model with endogenous population, in which fertility and current consumption / saving decisions are determined jointly. The model abstracts from the age and sex structure of the endogenous population, but its comparative statics and dynamics suggest a recovery pattern taking several generations. The increase in the capital-labor ratio would raise the wage rate (or level of wellbeing by any other name), as noted above. Consumption would be temporarily high, which would depress saving, as would be fertility if the income effect dominated a price effect in child-raising, as seems likely in a situation from antiquity in which there was limited scope for investment in children's human capital. The higher fertility and higher consumption are eliminated gradually and the economy returns to its preshock capital-labor ratio, interest and wage rates and fertility. However, because of the period of reduced saving, the absolute level of the capital stock does not return to its preshock level, and the temporary surge in fertility does not fully restore the preshock population before the preshock steady-state values of the endogenous variables are reached. The capital stock, once it resumes its trend with a growing population, remains below what it would have been without the population shock. Clearly, other changes could occur that would propel a recovered but stunted society beyond its revived steady-state configuration – after all, what I have

just described is simply a model, which has the virtue of suggesting how changes could have occurred, but it is a blind device and does not contain information beyond the behavioral and technological relationships it specifies – the rest is history rather than theory, but the theory suggests that other events must have occurred for the recovering society to eventually surpass its preshock circumstances.

Beyond the undifferentiated population in the Barro-Becker model, the age and sex composition of the surviving population could have been seriously disturbed, which would have affected fertility responses to reduced subsistence constraints (assuming such constraints were loosened by the reductions in the labor-capital and labor-land ratios). Analysis suggests that, assuming a population was in equilibrium (that is, its age structure was unchanging) prior to the shock, nearly a century could pass before it would return half-way back to an equilibrium size, given typical magnitudes of the response of the subsistence level to the population level and the response of population growth to the subsistence level (Lee 1993, 9–13; 1997, 1075). Further, with consumption and saving both affected by changes in the age composition of the surviving population (noting that saving is just production minus consumption and child-raising expenses), the pre-existing consumption and saving propensities of age groups could change as a consequence of altered conditions.

To focus the preceding series of ideas on a concrete event, consider the disintegration of Mycenaean civilization. Apparently its decline and disappearance was a considerably more complicated phenomenon than a quick reduction in either the capital stock or the population, despite the temporally concentrated remains of fire destructions. Certainly (or most probably) the fixed infrastructural capital stock declined but no reasons have been offered as to why the family-level capital stock involved primarily in agricultural production would have been affected. Why population, mostly rural at the time, declined, needs explanation outside the destruction or deterioration of the physical infrastructure observable as destroyed palaces. Did the palace infrastructure include an

institutional framework, supporting marketing of agricultural products, implying that individual rural families were integrated into markets that disappeared and caused havoc on farmsteads? Whether apparent population declines were the results of rational choices by individuals expecting a future worse than the present or the past or the result of exogenous events that killed off a considerable portion of the population has not, to my knowledge, been addressed. Another way of looking at the issue is to posit that extensive capital was destroyed, blurring the distinction between public, infrastructural capital and private capital, and that with no technological change, population had to decrease to recover the optimal capital-labor ratio. That would have taken a number of generations, and is a phenomenon that forward-looking economic theory has not put under its microscope.

14.6.4 Summary

Many influences can raise output or consumption per worker or per capita. Most make a contribution that does not continue: relative price changes, institutional improvements, exogenous changes in factor proportions. A once-and-for-all technological change would have a similar effect: its impact on output per worker would be reached asymptotically and thereafter the innovation would have no further effect. The effect would be permanent, but not a source of continuing growth in output per worker. A continuing flow of technological improvements that affected both physical and human capital would have conferred a continuing growth of output per worker, with whatever adjustments in savings and investment were required to keep the capital-labor ratio at its optimal level.

Observations of ancient standards of living may be easier to acquire than to evaluate. Changes in physical stature and other characteristics of skeletal remains may not bear one-to-one relationships to changes in real income. Apparent changes in house remains may reflect systematic gaps in evidence or comparisons between different parts of urbanized areas. When changes in some of these indicators can be identified in the physical record, it would be useful to tabulate

the array of possible circumstantial changes that could have led to the observations, from price changes to institutional changes to technological changes, episodic or continuing.

14.7 Suggestions for Using the Material of this Chapter

Ancient historians and philologists have published conclusions, tentative and less so, on economic growth at different times and places in antiquity. They have access to considerable written evidence on technology and to some extent on human capital reflected in the ancient analyses of agricultural and other technologies. Reliable evidence on consumption levels, or even changes in consumption levels over time, has been more elusive, so direct observations of economic growth have been more difficult to generate. The following thoughts will not sharpen quantification of ancient economic growth but are offered in the spirit of modestly heightening awareness of some major concepts of growth theory when thinking about various types of ancient evidence. In some instances, inferences of either growth or improved living standards (achieved of course through growth) may be sharpened, while in others they may be qualified because the evidence may be the results of other forces than growth.

14.7.1 Evidence of growth

Archaeologists, ancient historians, and philologists all look for evidence of economic growth in the periods and regions they study. The kinds of evidence found or sought fall into two categories, those representing the production capacity of an ancient society and those representing what individuals consumed. It is worth distinguishing between these two kinds of evidence.

The production side

Three phenomena could have increased the productive capacity of an ancient economy: a higher capital-labor ratio within a given technology regime, improved technology, and increased human capital. An archaeologist might, on a

lucky day on a well-preserved site, distinguish the effective amount of capital equipment or structure remains that could be associated with a relatively constant number of inhabitants. “Effective” is a critical adjective. Within the same general technology embodied in the remains of some kind of capital equipment, if higher quality equipment were found at a later date than lower quality equipment, with the quality differences being capable of delivering greater services from the piece of equipment, it might be reasonable to infer a higher capital-labor ratio at the later date. That should increase productivity. If differences in capital-labor ratios at different dates in the same locale could be identified in archaeological remains, or even in textual evidence, they would suggest differences in levels of productivity, not in continuing growth, which would also require continuing technological advances.

Irrigation technology may be a particular place to look for evidence of technological change, but that particular capital stock brings up the issue of private versus public capital goods. The contribution of public infrastructure to economic growth has been a heated subject in economics during the past several decades, reflecting as much late twentieth-century C.E. political ideologies as positive economics (archaeologists and ancient historians aren’t the only scholars susceptible to the disease), but differences between how variable and fixed capital stocks make their contributions to outputs should be kept in mind, since those characteristics will affect decisions regarding how much of each to supply. The degree of publicness or privateness of irrigation capital may be a matter of time and place, but scholarly tradition typically places it closer to the public end of the scale in Egyptian and ancient Near Eastern settings. Highways, of course, are another component of public capital. Technological change in both types of public capital stock may differ from that in private capital stock. Governments may have been able to afford experimentation (ancient R&D by whatever name)²⁵ as well as greater resources to implement new techniques and less concern over escape of the benefits of a new design to people not directly paying for it – a major concern in private innovation.

Identifying ancient technologies and technological change has long been a forte of archaeologists, as they have uncovered changes in kiln technologies that allowed finer control of heat and oxygen during firing of ceramics or in metallurgical technologies that offered greater heat control or superior ability to alloy metals and a host of others.²⁶ These technological developments, some of which warrant the term revolution, gradually raised the productive capacities of the societies that developed them by allowing things to be done that could not be done before. Whether those technological advances also contributed to discernible increases in per capita consumption is another matter. For millennia, population growth appears to have absorbed increases in productive capacity, keeping large proportions of ancient populations’ consumption levels more or less the same over long periods. Accordingly, attempting to identify what must have been economic growth by identifying technological improvements must deal with the additional variable of endogenous population growth, and the inference of no impact on living standards could be drawn, in one sense correctly, but in another sense with an additional variable or set of relationships intervening between the observations of technology level and consumption level.

Human capital would have been very important but also is very difficult to identify in archaeological remains. However, it may be expected that a people living in the same place for a number of generations will learn about the many contributors to agricultural capacity of the place, which should have led to increased agricultural production. A group’s continued occupancy of a place for a multigenerational period should lead to increased productivity.

Fruits of growth

Increased consumption, in many forms, is the result of increased productive capacity. Archaeologists, ancient historians, and philologists long have reported direct or indirect indicators of increased consumption. First, more robust skeletal stature probably reflects improved nutrition throughout a lifetime, subject to the caveats noted

above. That said, evidence of poorer health in urban settings may reflect conditions of crowding rather than constriction of consumption.

Second, increased urbanization is an indicator of an increase in the amount of food beyond family needs that a farming technology can produce and that transportation technology can deliver. The extent to which growth in urbanization reflects price effects without technological change or technological changes driving prices may remain an open question. Thus, a society may be capable of supporting a larger proportion of people not producing food than it actually does for some period of time, because the demand for what people might produce off farm is low. If the demand for those nonagricultural products increases, people can shift out of agriculture without technological advances in agriculture. On the other hand, it could be possible for advances in hydraulic technology in a dry area to free up agricultural workers by increasing output per unit of land. Different causes could generate similar changes in urbanization.

Archaeologists and their art-historical relatives have long codified qualitative advances in ceramics: durability, proportion of fine ware, general qualitative differences in ceramics – fineness of biscuit, accuracy of firing and oxygen control to govern colors, and various dimensions of attention to artistry. Artistry itself requires more skill, which must be acquired, and care in execution, both of which involve time, which as everyone now knows, equals money. A simplification of course, but higher quality ceramics, both fine wares and utilitarian wares, are evidence of increased consumption and consequently of economic growth – or at least a higher living standard at some period relative to an earlier period.

Scholars of antiquity have long associated larger dwelling spaces – houses – with higher living standards. Homes are a major capital item of most individuals, so a larger amount of house should be a good indicator of wealth levels at different times and places. That said, allowance must be made for dampening effects of higher land prices in high-density urban settings. Comparison of comparable conditions is important for inferences of income growth or of income differences between different sites based on house size – and of course, quality of construction.

14.7.2 Sectoral structure

The concept of sectors of an economy based on the activities conducted or on the locations of the activities can be useful to highlight variations in what were all essentially agricultural societies. Agriculture versus manufacturing, even if proportions of output elude even rough assessment, can be a useful dichotomy for judging the productivity of agriculture. Similarly with the proportion of a population living in a city (which might look like a big town to us) versus that living in the countryside, either on dispersed farmsteads or in small farm villages: most of the people living in the city didn't make much, if any, of their living from agricultural production, so the agriculturalists must have been capable of producing enough food to feed them, in exchange, of course, for the good things that the city dwellers produced. Again, quantification will be elusive, if not impossible, but it may be possible to draw useful judgments about probable increases in living standards at one time compared to an earlier or subsequent time from information on proportions of population in different sectors.

References

-
- Aghion, Philippe, and Peter Howitt. 1998. *Endogenous Growth Theory*. Cambridge MA: MIT Press.
- Arnold, Dieter. 1991. *Building in Egypt; Pharaonic Stone Masonry*. Oxford: Oxford University Press.
- Barro, Robert J., and Gary S. Becker. 1989. "Fertility Choice in a Model of Economic Growth." *Econometrica* 57: 481–501.
- Becker, Gary S., and H. Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy* 81: S279–S288.
- Becker, Gary S., and Kevin M. Murphy. 1992. "The Division of Labor, Coordination Costs, and Knowledge." *Quarterly Journal of Economics* 107: 1137–1160.

- Bureau of the Census, U.S. Department of Commerce. 1975. *Abstract of U.S. Historical Statistics*. Washington, D.C.: U.S. Government Printing Office.
- Chenery, Hollis B. 1979. *Structural Change and Development Policy*. New York: Oxford University Press.
- Clarke, Somers, and R. Engelbach. 1930. *Ancient Egyptian Masonry*. Oxford: Oxford University Press.
- Cuomo, S. 2007. *Technology and Culture in Greek and Roman Antiquity*. Cambridge: Cambridge University Press.
- Denison, Edward F., assisted by Jean-Pierre Poulhier. 1967. *Why Growth Rates Differ. Postwar Experience in Nine Western Countries*. Washington, D.C.: Brookings Institution.
- Denison, Edward F. 1974. *Accounting for United States Economic Growth 1929–1969*. Washington, D.C.: Brookings Institution.
- Dixit, A.K. 1976. *The Theory of Equilibrium Growth*. Oxford: Oxford University Press.
- Easterlin, Richard A. 1966a. "On the Relation of Economic Factors to Recent and Projected Fertility Changes." *Demography* 3: 131–153.
- Easterlin, Richard A. 1966b. "Economic-Demographic Interactions and Long Swings in Economic Growth." *American Economic Review* 56: 1063–1104.
- Easterlin, Richard A. 1974. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." In *Nations and Households in Economic Growth; Essays in Honor of Moses Abramovitz*, edited by Paul A. David and Melvin W. Reder. New York: Academic Press, pp. 89–125.
- Easterlin, Richard A. 1978. "What Will 1984 Be Like? Socioeconomic Implications of Recent Twists in Age Structure." *Demography* 15: 397–432.
- Frey, Bruno, and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40: 402–435.
- Giannecchini, Monica, and Jacopo Moggi-Cecchi. 2008. "Stature in Archaeological Samples from Central Italy: Methodological Issues and Diachronic Changes." *American Journal of Physical Anthropology* 135: 284–292.
- Goldsmith, Raymond W. 1984. "An Estimate of the Size and Structure of the National Product of the Early Roman Empire." *Review of Income and Wealth* 30: 263–288.
- Hitchner, Bruce. 2005. "The Advantages of Wealth and Luxury: The Case for Economic Growth in the Roman Empire." In *The Ancient Economy; Evidence and Models*, edited by J.G. Manning and Ian Morris. Stanford CA: Stanford University Press, pp. 207–222.
- Ho, Samuel P.S. 1978. *Economic Development of Taiwan, 1860–1970*. New Haven CT: Yale University Press.
- Holleran, Claire. 2012. *Shopping in Ancient Rome; The Retail Trade in the Late Republic and the Principate*. Oxford: Oxford University Press.
- Hopkins, Keith. 1995/96. "Rome, Taxes, Rents and Trade." *Kodai: Journal of Ancient History* VI/VII: 41–75. Reprinted in *The Ancient Economy*, edited by Walter Scheidel and Sitta von Reden. London: Routledge, pp. 190–230.
- Hopkins, Keith. 2002. "Rome, Taxes, Rent and Trade." In *The Ancient Economy*, edited by Walter Scheidel and Sitta von Reden. New York: Routledge, pp. 190–230.
- Jones, Charles I. 1998. *Introduction to Economic Growth*. New York: Norton.
- Katzman, Martin T. 1977. *Cities and Frontiers in Brazil: Regional Dimensions of Economic Development*. Cambridge MA: Harvard University Press.
- Kelder, Jorrit M. 2010. *The Kingdom of Mycenae; A Great Kingdom in the Late Bronze Age Aegean*. Bethesda MD: CDL Press.
- Kim, Dae Young, and John E. Sloboda. 1981. "Migration and Korean Development." In *Economic Development, Population Policy, and Demographic Transition in the Republic of Korea*, edited by Robert Repetto, Tae Hwan Kwon, Son-Ung Kim and Dae Young Kim. Cambridge MA: Harvard University Press, pp. 36–138.
- Komlos, John, and Marieluise Baur. 2004. "From the Tallest to (One of the) Fattest: The Enigmatic Fate of the American Population in the 20th Century." *Economics and Human Biology* 2: 57–74.
- Küpper, Michael. 1996. *Mykenische Architektur: Material, Bearbeitungstechnik, Konstruktion und Erscheinungsbild*, Internationale Archäologie Band 25. Espelkamp: Marie Leidorf.
- Kuznets, Paul W. 1977. *Economic Growth and Structure in the Republic of Korea*. New Haven CT: Yale University Press.
- Kuznets, Simon. 1966. *Modern Economic Growth; Rate, Structure and Spread*. New Haven CT: Yale University Press.
- Lancaster, Lynne. 2008. "Roman Engineering and Construction." In *The Oxford Handbook of Engineering and Technology in the Classical World*, edited by John Peter Oleson. Oxford: Oxford University Press, pp. 256–284.
- Lebergott, Stanley. 1976. *The American Economy: Income, Wealth, and Want*. Princeton NJ: Princeton University Press.
- Lee, Ronald D. 1987. "Population Dynamics of Humans and Other Animals." *Demography* 24: 443–465.

- Lee, Ronald D. 1993. "Accidental and Systematic Change in Population History: Homeostasis in a Stochastic Setting." *Explorations in Economic History* 30: 1–30.
- Lee, Ronald D. 1997. "Population Dynamics: Equilibrium, Disequilibrium, and Consequences of Fluctuations." In *Handbook of Population and Family Economics, Volume 1B*, edited by Mark R. Rosenzweig and Oded Stark. Amsterdam: Elsevier, pp. 1063–1114.
- Levy, David. 1984. "Testing Stigler's Interpretation of 'The Division of Labor Is Limited by the Extent of the Market'." *Journal of Industrial Economics* 32: 377–389.
- Lewis, Michael J.T. 1997. *Millstone and Hammer: The Origins of Water Power*. Hull: University of Hull Centre for Southeast Asian Studies.
- Locay, Luis. 1990. "Economic Development and the Division of Production between Households and Markets." *Journal of Political Economy* 98: 965–982.
- Lucas, A., and J.R. Harris. 1962. *Ancient Egyptian Materials and Industries*, 4th edn. London: Edward Arnold.
- Lucas, Robert E., Jr. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22: 3–42.
- Macunovich, Diane J. 1998. "Fertility and the Easterlin Hypothesis: An Assessment of the Literature." *Journal of Population Economics* 11: 53–111.
- Maddison, Angus. 1971. *Class Structure and Economic Growth: India and Pakistan since the Moghuls*. New York: Norton.
- Maddison, Angus. 2007. *Contours of the World Economy, 1–2030 AD; Essays in Macro-Economic History*. Oxford: Oxford University Press.
- Mamalakis, Markos J. 1975. *The Growth and Structure of the Chilean Economy from Independence to Allende*. New Haven CT: Yale University Press.
- Mitchell, B.R. 1982. *International Historical Statistics; Africa and Asia*. New York: New York University Press.
- Mitchell, B.R. 1983. *International Historical Statistics; The Americas and Australasia*. Detroit: Gale Research.
- Moorey, P.R.S. 1994. *Ancient Mesopotamian Materials and Industries; The Archaeological Evidence*. Oxford: Clarendon Press.
- Morris, Ian. 2004. "Economic Growth in Ancient Greece." *Journal of Institutional and Theoretical Economics* 160: 709–742.
- Morris, Ian. 2008. "Early Iron Age Greece." In *The Cambridge Economic History of the Greco-Roman World*, edited by Walter Scheidel, Ian Morris, and Richard Saller. Cambridge: Cambridge University Press, pp. 211–241.
- Nerlove, Marc, and Kakshmi K. Raut. 1997. "Growth Models with Endogenous Population: A General Framework." In *Handbook of Population and Family Economics, Volume 1B*, edited by Mark R. Rosenzweig and Oded Stark. Amsterdam: Elsevier, pp. 1117–1174.
- Nicholson, Paul T., and Ian Shaw, eds. 2000. *Ancient Egyptian Materials and Technology*. Cambridge: Cambridge University Press.
- North, Douglass C. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.
- Oleson, John Peter, ed. 2008. *The Oxford Handbook of Engineering and Technology in the Classical World*. Oxford: Oxford University Press.
- Oleson, John Peter, Christopher Brandon, and Robert L. Hohlfelder. "Technology, Innovation, and Trade: Research into the Engineering Characteristics of Roman Maritime Concrete." In *Maritime Archaeology and Ancient Trade in the Mediterranean*, edited by Damian Robinson and Andrew Wilson. Oxford: Oxford Centre for Maritime Archaeology, Institute of Archaeology, pp. 107–119.
- Pampel, Fred C., and H. Elizabeth Peters. 1995. "The Easterlin Effect." *Annual Review of Sociology* 21: 163–194.
- Pilkington, Nathan. 2013. "Growing Up Roman: Infant Mortality and Reproductive Development." *Journal of Interdisciplinary History* 44: 1–35.
- Potts, D.T. 1997. *Mesopotamian Civilization; The Material Foundations*. Ithaca NY: Cornell University Press.
- Romer, Paul. 1987. "Crazy Explanations for the Productivity Slowdown." In *NBER Macroeconomics Annual*, edited by Stanley Fischer. Cambridge MA: MIT Press, pp. 163–202.
- Rosen, Sherwin. 1978. "Substitution and Division of Labour." *Economica* NS 45: 235–250.
- Rosen, Sherwin. 1983. "Specialization and Human Capital." *Journal of Labor Economics* 1: 43–49.
- Saller, Richard. 2002. "Framing the Debate over Growth in the Ancient Economy." In *The Ancient Economy*, edited by Walter Scheidel and Sitta von Reden. New York: Routledge, pp. 251–269.
- Saller, Richard. 2005. "Framing the Debate over Growth in the Ancient Economy." In *The Ancient Economy; Evidence and Models*, edited by J.G. Manning and Ian Morris. Stanford CA: Stanford University Press, pp. 223–228.
- Scheidel, Walter. 2009. "In Search of Roman Economic Growth." *Journal of Roman Archaeology* 22: 46–70.

- Shaw, Joseph W. 1971. *Minoan Architecture: Materials and Techniques*, *Annuario della Scuola Archeologica di Atene* 49, N.S. 33. Rome: L'Erma di Bretschneider.
- Shaw, Joseph W. 2009. *Minoan Architecture: Materials and Techniques*. Studi di Archaeologia Cretesi VII. Padua: Bottega D'Erasmio.
- Silver, Morris. 2007. "Roman Economic Growth and Living Standards: Perceptions versus Evidence." *Ancient Society* 37: 191–252.
- Stevenson, Betsey and Justin Wolfers. 2008. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." *Brookings Papers on Economic Activity*, Spring: 1–87.
- Stigler, George J. 1951. "The Division of Labor Is Limited by the Extent of the Market," *Journal of Political Economy* 59: 185–193.
- Voutsaki, Sofia. 2010. "Mainland Greece." In *The Oxford Handbook of the Bronze Age Aegean*, edited by Eric H. Cline. Oxford: Oxford University Press, pp. 99–112.
- Wilkie James W., and Adam Perkal, eds. 1984. *Statistical Abstract of Latin America* Vol. 23. Los Angeles CA: UCLA Latin American Center.
- Wilson, Andrew. 2002. "Machines, Power and the Ancient Economy." *Journal of Roman Studies* 92: 1–32.
- Wilson, Andrew. 2009. "Indicators for Roman Economic Growth: A Response to Walter Scheidel." *Journal of Roman Archaeology* 22: 47–82.

Suggested Readings

- Aghion, Philippe, and Peter Howitt. 2009. *The Economics of Growth*. Cambridge MA: MIT Press. Introduction and Chapters 1–2.
- Barro, Robert J., and Xavier Sala-i-Martin. 2004. *Economic Growth*, 2nd edn. Cambridge MA: MIT Press. Chapters 1–2.
- Jones, Charles I. 2002. *Introduction to Economic Growth*, 2nd edn. New York: Norton. Chapters 1, 4–5.
- Weil, David N. 2009. *Economic Growth*, 2nd edn. Boston MA: Addison-Wesley. Chapters 1–4.

Notes

- 1 Scheidel (2009) and Wilson (2009) have debated the interpretation of several popular indicators of the pace of economic activity but more from the perspective of archaeological biases than economic implications.
- 2 Maddison updated Raymond Goldsmith's (1984) estimate of Roman gross domestic product (GDP) per capita for 14 C.E. and found the grain-equivalent version of his estimate to be 42.3% of the estimated GDP per capita for England and Wales in 1688 (Maddison 2007, 52). From 42% to 100% sounds like a lot of growth but over that time period it amounts to an average annual growth rate of 0.051%, barely perceptible. Granted, this is an average that includes a likely retrenchment of income per capita during the later Empire and early medieval period and thus leaves some room for one or more periods of higher than average growth, but 0.051% is a low bar. However, the Late Imperial and Early Medieval depression of incomes may have affected the average income far more than the median income since people toward the large, lower tail of the considerably skewed income distribution of the times did not have that far they could fall before they starved

to death – granted that adjustments such as shrinkage of stature and even poorer health were possible mediators. That said, Giannellini and Moggi-Cecchi (2008) report a reduction in the height of Roman skeletal evidence from the Italian Iron Age to the Roman period – fifth century B.C.E. to fifth century C.E. – and a subsequent increase in height following the Roman period. They are unable to narrow their findings to shorter periods. This reasoning from later consumption levels to previous growth is of the sort that Hopkins (1995/96; 2002, 194, n. 4) called his "fourth framing principle" – look for the implications of one contention on other factors that are known.

Maddison also estimated per capita GDP for a number of regions in the Roman Empire in 300 B.C.E. and 14 C.E. (57, Table 1.14). Again, the average annual growth rates implied are quite low: the highest is for peninsular Italy, at 0.223%; the average growth rate for all the regions combined, including peninsular Italy, is 0.067%, which, considering the scope for error in the estimates for the beginning and ending years, is effectively no different from the average growth rate between 14 C.E. and 1688.

- 3 Morris (2004, 2008) has cited possible indicators of economic growth in Greece between the Middle Geometric Period (ca. 800 B.C.E.) and the early Hellenistic Period (ca. 300 B.C.E.) as well as possible contributing causes, and has suggested a period of comparable economic growth on Minoan Crete between 2000 and 1500 and B.C.E.
- 4 On Roman economic growth, see Hitchner (2005), Saller (2002, 2005), Silver (2007), Scheidel (2009), and Wilson (2009). Holleran (2012, 39–43) reviews the evidence and offers the summary judgment that the case for noticeable growth is not persuasive. Saller (2002, Fig. 12.2, p. 260), suggests an increase in Roman productivity (GDP per capita) by some 25% between 200 B.C.E. and ca. 75 C.E., with a somewhat more rapid decline back to 200 B.C.E. levels by 300 C.E., citing Hopkins (2002), but without the help of the graphical presentation.
- 5 The following presentation is adapted from Jones (1998, Chapter 2) and Dixit (1976, Chapter 3).
- 6 For those who want to follow the derivation of this version of the accumulation equation, begin with the definition of k : $k \equiv K/L$, which in logarithms is $\log k = \log K - \log L$. Then $\dot{k}/k = \dot{K}/K - \dot{L}/L$, where each term is a rate of change. Going back to the original accumulation equation, $\dot{K} = sQ - dK$, or $\dot{K}/K = sQ/K - d$. Now, substitute this rearrangement of the accumulation equation into the logarithmic expression for the rate of change in the capital/labor ratio to get $\dot{k}/k = sQ/K - d - \dot{L}/L = sQ/K - d - n$, where n is the population (labor) growth rate. Next, in the first term on the right-hand side, divide both output and capital by labor to get $\dot{k}/k = sq/k - n - d$, from which multiplication of both sides by k gives the per capita accumulation.
- 7 You can see this by setting $\dot{K} = 0$ in $\dot{k}/k = \dot{K}/K - \dot{L}/L$.
- 8 Recall that $sy = sk^\alpha$.
- 9 A third type of mechanism relies on a particular mathematical property, naturally with efforts to justify it: constant returns to scale in the application of capital, called the AK model (Romer 1987). This contrasts with the usual, neoclassical property of diminishing returns to scale in each separate factor, independently of whether a process experiences constant, increasing, or decreasing returns to scale in all factors applied simultaneously. With constant returns to scale in capital, the curved investment line of the Solow diagram, the sq curve, becomes a straight line. If the sq line lies above the investment requirements line – the $(n + d)k$ line – investment always will generate more capital than is required to maintain the capital / labor ratio at the same level, so k will increase literally without bound, as will production per capita, q .
- 10 For text treatments of the incorporation of R&D into growth models, see Jones (1998, Chapters 5 and 8) and Aghion and Howitt (1998, Chapters 3 and 6).
- 11 This model is based on Lucas (1988), as modified in Jones (1998, 48–50).
- 12 For this section, I have relied on the following: Stigler (1951), Rosen (1978, 1983), Levy (1984), Locay (1990), and Becker and Murphy (1992).
- 13 International trade is a tricky example, if a natural one, of a source of market expansion intended to explain increased productivity resulting from increased specialization attendant upon a larger market. If the opening of trade benefits the people of both trading countries, that means it will increase their incomes, their production per worker and consumption per worker, with no change in either specialization of firms or factors or technological change to reflect increased specialization in the capital stock. Thus attributing all of the increase in observed productivity (should we be so lucky as to be able to observe it in evidence from 2000 to 3000 years ago) following the opening of trade to increased specialization would be an overstatement. Disentangling the two effects empirically would be challenging.
- 14 Of course, good agricultural land can be scarce in rural areas, leaving considerable competition for living space, which may be constrained to more rugged terrain. In such a case, the more accessible sites in the terrain unsuited for agriculture may go for a premium, and we have competition for both space and locations analogous to what occurs in cities.
- 15 If our sectoral distinction had been urban / rural, we might have wanted to distinguish between possibly differing preference systems of city dwellers and people in the countryside. We would need a separate demand relationship for each category of worker in that case. I have made no such distinction between the preferences of agricultural and industrial workers, although such might be justified in some cases.
- 16 We must be careful of what our language implies here. Clearly we, as students and modelers of such an economy, are not using any of these resources. However, when we think about how the agents in the economy themselves go about their

- allocations, we want to be sure that, in our role, we recognize the constraints they face(d). They can't use more than they have, and they will use all they do have, and good modeling should account for these "adding-up" conditions.
- 17 "High," of course, is relative. A stream of benefits, such as the harvest on an acre planted in wheat, that would have been thought "high" or "good" in the second millennium B.C.E. in northern Syria would be thought unacceptably low in many parts of the world in recent years. We should consider "high" relative to the stream of benefits that could be sustained from a given resource with a given technology, which is roughly how much of the potential people are getting out of what they possess, including their knowledge.
 - 18 These two terms have acquired relatively short-term, business-cycle meanings in contemporary expression, whereas I have in mind periods that can last several generations to centuries, occupying all or the greater part of some relative chronological periods in some regions.
 - 19 There is recent economic research into the relationship(s) between how much people consume and how happy they are but it's in its youth, and over the range of increase in material consumption we're dealing with in these cases it is reasonable to associate increases in utility (the economic code word for wellbeing) with the increases in output per worker. The initial inquiry into the correspondence between measures of income or output such as GDP and happiness is Easterlin (1974), introducing the Easterlin paradox, whereby within-country surveys of personal happiness corresponded closely to income, but in inter-country comparisons, happiness and measures of income bore no relationship. Frey and Stutzer (2002) provide a useful progress report. Stevenson and Wolpers provide within- and cross-country empirical evidence that conventional, if imperfect measures of wellbeing such as GDP per capita accord well with results of individual happiness surveys across as well as within countries, suggesting the Easterlin paradox is a product of data and statistical methodologies. At the risk of sounding reductionist, the issue to a considerable extent has devolved in recent years into an empirical discovery of the decreasing marginal utility of cash income, which is not controversial.
 - 20 Education, which increases the quality, or productivity, of labor, accounts for roughly 10% of observed growth recently in industrial countries, and improvements in the quality of the capital stock a comparable proportion (Denison 1967, Chapter 21; 1974, Chapter 9). The former effect is growth in human capital and the latter is part of technological change. The extent to which increases in literacy, and possibly numeracy, in antiquity may have raised labor productivity is an open question. The increase in capital quality could be rolled into the one-quarter to one-third of output growth that eluded assignment to specific causes in growth accounting and has been dubbed either total factor productivity growth or technological change.
 - 21 This characterization parallels Lucas's distinction between "level effects" and "growth effects" in the evaluation of institutional changes in contemporary developing countries which various authors have characterized as limits to growth (Lucas 1988, 12).
 - 22 Useful summaries of work through the mid-1990s are Lee (1997) and Nerlove and Raut (1997).
 - 23 According to Lee (1987), Malthusian-style controls on population, such as deterioration of wages or other measures of wellbeing at larger population sizes, are not particularly important in short- and medium-run fluctuations in human populations, although they are important homeostatic mechanisms in the long run and form important adjustment mechanisms in recovery from exogenous nondensity-related shocks, such as considered here. Ancient populations simply did not grow rapidly enough to crash noticeably – either to them or to us – into Malthusian walls, so testing of Malthusian predictions requires considerable subtlety.
 - 24 Pampel and Peters (1995) and Macunovich (1998) provide useful overviews of a broad body of research sparked by Easterlin's cohort-size ideas.
 - 25 Lancaster (2008, 260) characterizes the third-century B.C.E. Roman developments with concrete as "beginning to experiment," whether the choice of words as suggesting systematic investigations is intended or not. She proceeds (260–262) to describe the technological improvements in concrete over the following two centuries and their impacts on construction costs. On the same subject of Roman concrete, particularly for maritime applications, Oleson *et al.* (2011) note the very similar harbor construction technologies in Italy and the eastern Mediterranean from the late second-century B.C.E. through the first-century C.E. and suggest either "the movement of engineers" or "the circulation of subliterate technical manuals of which no trace has survived" (other

than Vitruvius) as being responsible for the close similarities. Either mechanism would indicate some kind of systematization of this engineering knowledge as well as experimentation – after all, who would recommend an untried technique in such highly visible constructions as the Temple of Magna Mater or a major harbor installation?

- 26 While many studies trace apparent developments in technologies in antiquity as part of projects with other aims, much basic material on various ancient technologies has been made available over the years in works devoted to those topics, a number of them standard references by now. For example, for Egyptian industrial technologies, there are the classic Lucas and Harris (1962), supplemented more recently by Nicholson and

Shaw (2000), while the construction industry has been served by the long-standing Clarke and Engelbach (1930), also supplemented recently by Arnold (1991). For Mesopotamia, there are Moorey (1994) and Potts (1997). A recent reference work on Greek and Roman technologies is Oleson (2008), which cites the numerous ancient writers on industrial processes and fixed and mobile engineering, while Shaw (1971; 2009) establishes a foundation for building technology on Minoan Crete, and Küpper (1996) establishes a similar base for Mycenaean Greece. A list such as this could be extended nearly indefinitely, but these old standbys and newer contributions contain an extensive base of information for the analysis of technological change.

Index

- ability-to-pay principle 149–150
- absolute efficiency 19
- accumulation of capital 263–264, 276–279, 282–283, 396, 538–541
- Achaemenid Babylonia 25–26
- adaptive expectations 246–249
- adding-up condition 4, 321–322
- adjustment costs 287
- ad valorem* tax 151, 153, 155
- adverse selection 236, 239–240, 254
- AFC *see* average fixed cost
- ageing of capital 266–268
- agglomeration economies 476–477
- aggregate demand 65–66, 77, 88–89, 97, 460–461
- aggregate growth 535
- aggregate income 321–322
- aggregate savings 294
- agriculture
 - capital 259–261
 - cities 485–488
 - cobweb model of supply 248, 250–251
 - consumption 90–93
 - cost 42–43, 50–51
 - growth 537, 546–551, 558–561
 - production 15–16, 19–25
 - public economics 149, 153, 188–189
 - see also* household production theory; land use
- Allison, Penelope M. 95–96
- all-or-nothing work offers 362–363
- altruism 405–406
- AP *see* average product
- Appianus estate 106
- applicability of economic theory 1–2, 4–6
- aqueducts 504–505
- arithmetic mean 207–209
- Arrow impossibility theorem 183–184, 185
- asset transformation 323–324, 328–329
- asymmetric information 218, 235–242
 - adverse selection 236, 239–240, 254
 - moral hazard 236, 237–239
 - principal–agent relationships 236, 237–239, 253–254
 - production 45
 - signaling and screening 236, 240–242
- auctions 184–185
- AVC *see* average variable cost
- average cost (AC)
 - capital 271
 - competition 109, 112–113
 - growth 545
 - land use 456–457
 - public economics 142, 176
- average fixed cost (AFC) 31
- average product (AP) 11–12
- average variable cost (AVC) 31–32
- Averch–Johnson effect 193–194
- Babylonia 25–26, 319
- backstop technology 519–520
- balance sheets 323–326, 329
- banking
 - asset transformation 323–324, 328–329
 - balance sheets 323–326, 329
 - banking firms 328–332
 - bank money 305, 324–328
 - bimetallism 337

- banking (*continued*)
 - capital 262
 - creation of money 305, 323–328
 - currency 325–328, 335–336
 - demand deposits 323–324, 328
 - demand for money 310, 314
 - financial intermediation 332–335
 - foreign exchange 335–336, 345
 - measurement of money 310
 - monetary policy 343–345
 - neoclassical quantity theory 314
 - seigniorage 335–337, 343–345
 - supply of money 318–319
 - theory of 252
 - withdrawal rates 324–325, 329
- bankruptcy 329–330
- Barro–Becker model 558
- Bayesian probability theory 216, 218–223, 254
- Becker–Lewis model 557–558
- behavioral uncertainty 235–246
 - asymmetric information 235–242
 - competitive behavior 253–254
 - strategic behavior 242–246
- benefits approach to taxation 149–150
- Benthamite social welfare function 150–151
- bid-rent function 447–450, 488–489
- bimetallism 337
- bonds 316–318
- bonuses 386
- book value 269
- Borda score 184
- bounded rationality 4
- budget
 - consumption 55, 57–58, 59–61, 74–75
 - general equilibrium 125, 129–130
 - labor 358, 411, 412–413, 417
 - public economics 160, 162, 179
- bureaucracy 195–196
- Cahill, Nicholas D. 95
- Cambridge cash balance 313–314, 315
- capital 258–300
 - accumulation 263–264, 276–279, 282–283, 396, 538–541
 - ageing of capital 266–268
 - capital deepening/widening 539
 - capital richness 283–284
 - cities 494–495, 498–502, 507–508
 - concepts and definitions 258–270
 - consumption and saving 290–294
 - demand for and supply of 279–283
 - durability 265–266, 267
 - embodiment 268
 - formation of 294–296
 - growth 537–542, 547–549, 556–560
 - human capital 259–260
 - interest rates 272–284
 - inventories 260–261, 262–263
 - investment 276–282, 284–287
 - labor 376–378, 383–384, 399–400, 537–542, 547–549, 556–560
 - labor theory of value 269–270
 - land use 467
 - maintenance 287–289
 - material capital 259–262
 - measurement of 268–269
 - mobility of 507–508
 - money and banking 262, 329
 - prices and values 264–265
 - production 262–263
 - public and private capital stock 261
 - quasi-rents 270–272
 - scrapping and replacement 289–290
 - stocks and flows 263–264, 279–283
 - as stuff 259–262
 - temporal aspects 265–268, 274–275, 290–291
 - theory of capital 276–284
 - use by firms 284–290
- capital asset pricing model (CAPM) 235
- capitalization of risk 234–235
- CAPM *see* capital asset pricing model
- CBA *see* cost–benefit analysis
- CBD *see* central business district
- cenaculae* 95
- central business district (CBD) 482
- centralization 139–140
- central place theory 441, 455–456, 457, 461–463, 492
- ceramics *see* pottery/ceramics
- certainty equivalence 210–213
- CES *see* constant elasticity of substitution
- CGEMs *see* computable general equilibrium models
- change-resistant beliefs 222
- characteristics model 79–81
- checking accounts 323–324, 328
- child care 352, 362–363
- child labor 407–414
- cities 472–515
 - activities and functions 473–474
 - ancient versus contemporary cities 472–475
 - bid-rent function 488–489
 - central place theory 492
 - characteristics of 473
 - classifications of 472–473, 478–479, 497–499
 - consumption 474, 478–480, 482–497
 - density gradients in ancient cities 491
 - economic base theory 478
 - economies of 475–479
 - endogenous centers 490–491
 - externalities 477
 - growth 547–551, 561
 - home workers 489–490
 - housing 473–474, 475, 479–488, 493–497, 506–507

- infrastructure 493–497, 502, 504–505
- land use 451–452
- location theory 482
- monocentric city model 483–489
- multiple resident categories 488–489
- production 474, 475–479, 486, 493–497
- public finance 503–507
- raising revenue 506–507
- scale economies in production 475–477, 494
- size distribution of cities 499–503
- spatialities 473, 482–492, 498–499
- systems of cities 492–503
- tradable and nontradable goods 477–478, 493–494
- types of production 477–479
- urbanization economies 476
- wage differentials across cities 491–492, 498–499, 501–502
- closed city model 482, 487–488
- Cobb–Douglas function
 - cities 493–494
 - consumption 60, 61–62
 - cost 33, 40
 - general equilibrium 123
 - growth 538–539, 541, 543–544, 557
 - natural resources 527
 - production 16–17
- coefficient of variation (CV) 208
- COL *see* cost-of-living
- commodity money 304, 318–323
- commodity standards 262
- commons 218, 523–524, 528–531
- commuting 362–363
- comparative statics 37
- compensated demand curve 62, 67–68
- compensating differentials 350, 387–391
- compensating variation 66–68
- competition
 - competitive equilibrium 106–108, 140
 - consumption 55, 80, 81–82
 - contestable markets 104, 112–113
 - cost 30, 39, 44
 - general equilibrium 134
 - industry structure 103–119
 - labor 252–253, 422–423
 - land use 448, 455, 466–467
 - monopolistic competition 80, 81–82, 104, 111–112, 114–115
 - monopoly 103–104, 108–110, 115–117, 134, 523–524
 - monopsony 104, 113–114
 - natural resources 520–521, 523–524
 - oligopoly 104, 110–111, 115–117
 - perfect competition 103–106, 114–115
 - production 19
 - public economics 192–193
- complements in production 379, 382–383
- complete markets 227–228
- computable general equilibrium models (CGEMs) 123, 134–136
- conditional efficiency 19
- conditional probability 219–221
- conflicts of interest 224
- constant elasticity of substitution (CES) 16–17, 60, 123
- constant returns to scale (CRS) 16–17, 39–40
- constrained maximization/optimization 36–37, 286
- construction
 - capital 261–262, 296–297
 - housing 480–481
 - production 21–22
 - public economics 188–191, 197
- consumption 55–102
 - aggregate demand 65–66, 77, 88–89, 97
 - budget 55, 57–58, 59–61, 74–75
 - capital 274, 276–279, 283–284, 290–294
 - characteristics model 79–81
 - cities 474, 478–480, 482–497
 - competition 55, 80, 81–82
 - concepts and definitions 55
 - cost 40, 60, 63, 72–73, 76–77, 83
 - demand 55, 60–66, 77–79, 88–99
 - discrete choice 77–78
 - durable goods 56, 75–79, 97
 - elasticity of demand 63–65, 79, 85, 93–94, 98
 - Engel curve 64–65
 - equivalent and compensating variation 66–67
 - estimation methods 65
 - expected utility 86–88
 - externalities 144–146
 - fixed prices 90–93
 - general equilibrium 125
 - growth 539–540, 547, 555
 - hedonic price model 79–80, 93
 - household production theory 78–79, 82–85
 - income 56, 63–65, 78–79, 83–85, 93–94
 - indifference curve 58–61, 64, 68, 73–75, 77–78
 - individual and aggregate savings 294
 - interest rates 283–284
 - intertemporal choice 56, 73–75
 - intertemporal utility maximization 290–291
 - labor 357–359, 365, 399, 410–411, 414–418
 - land use 441, 457–463
 - lifecycle savings model 292–294
 - long run and short run 65
 - markets 56, 62
 - monopolistic competition 80, 81–82
 - permanent income hypothesis 291–292
 - present and future consumption 276–279
 - price 56–58, 60–63, 66–77, 83–84, 90–93
 - price and consumption indexes 70–73
 - production 68–70, 78–79, 82–85

- consumption (*continued*)
 - public economics 140–146, 155, 158–164, 167–172, 186–190
 - rationality and irrational behavior 55, 57, 88–89
 - risk 60, 86–88, 212–215, 229–230
 - scale economies 80–81
 - substitution 59–60, 64, 78, 80, 84–85, 96–98
 - surplus 67–70, 91–92
 - tastes and preference 57, 75
 - transportation costs 69–70
 - uncertainty 56, 79, 228–230, 235
 - utility 58–60, 61–63, 66–69, 77–88
 - variety and differentiated goods 56, 79–82, 98
 - wellbeing 55, 59, 66–70
- contestable markets 104, 112–113
- contingent markets 225–230
- contractual agreements 384–391
 - basis of pay 385–387
 - compensating differentials 350, 387–391
 - concepts and definitions 350, 352, 384
 - information problems and incentives 384–385
 - sequencing of pay 387
- contractual theories of government 195
- cooperation 406
- corporation tax 166–167
- cost 29–54
 - accounting for apparent cost changes 49–50
 - allocation of factors across activities 43
 - asymmetric information 240–242
 - capital 286–290
 - cities 494–495, 498–499
 - concepts and definitions 29–30
 - constrained optimization problems 36–37
 - consumption 60, 63, 72–73, 76–77, 83
 - demand 30, 40–42, 47
 - Edgeworth–Bowley box diagram 51–52
 - elasticity of supply 38–40
 - entire economies 50–52
 - first- and second-order conditions 35–36, 37
 - growth 545
 - horizontal and vertical integration 44–45
 - industry structure 105, 108–109, 112–114
 - input demand curve 40–41
 - isoquant diagram 50–51
 - labor 375–376, 380–381, 391–393, 428
 - land use 442–447, 450–461, 463–467
 - long run and short run 29, 32–33
 - natural resources 517–522, 526–527, 530
 - opportunity cost 29, 42–43
 - optimization 30, 34–37
 - organization of production 43–46
 - price 35–36, 39, 43, 46–48, 52
 - primal–dual relationship 29
 - production 29–30, 33, 36–52
 - production possibilities frontier 50–52
 - profit maximization 30, 34–36, 44
 - public economics 142–145, 193–194
 - substitution 33, 34, 45
 - supply curves 35–40, 42
 - total cost curve 31–32
 - wages and rents 41–43
- cost–benefit analysis (CBA) 30
 - capital 266
 - labor 355, 392–393
 - public economics 142, 144–148, 186–191
- cost of capital 167
- cost-of-living (COL) index 72
- cost-push model of inflation 338
- cost shares 18, 23
- cotton 9, 17–18
- Cournot aggregation condition 64
- Cournot–Nash behavior 110–111
- covariance 231–235, 251
- credit money 304–305, 314
- cross-effects 59, 157–158, 182, 411
- cross-price elasticity of demand
 - consumption 63–65, 79, 94, 98
 - factors of production 41
 - labor 379, 381–383
 - public economics 167–168, 170–171
- crowding effect 529
- CRS *see* constant returns to scale
- currency
 - creation of money 324–328
 - foreign exchange 335–336
 - monetary policy 343–344
 - monetization prior to currency 303–304
- customary-price models 6
- CV *see* coefficient of variation
- Cyprus 117
- deadweight losses 155–158
- De Arboribus* (Columella) 15
- deben (Egypt) 5
- debt 173
- decentralization 196
- declining-cost industries 175–176
- deflation 307, 340–342
- deflators 71–72
- demand
 - aggregate demand 65–66, 77, 88–89, 97
 - capital 279–283, 288
 - competitive equilibrium 106–108
 - consumption 55, 60–66, 77–79, 88–99
 - contestable markets 112–113
 - cost 30, 40–42, 47
 - derived demand 41, 379–384, 426–427, 466
 - expectations 248–253
 - general equilibrium 124–127, 129–134
 - growth 548–549
 - housing 481–482, 494–496, 506–507
 - industry structure 104–114

- labor 3, 351–352, 375–384, 388–389, 401–403, 418–419, 426–427
- land use 441–442, 445–447, 460–461, 466
- money 301–302, 309–318, 321–322, 338–341, 345
- monopolistic competition 111–112
- monopoly 108–110
- monopsony 113–114
- natural resources 521, 523–524
- perfect competition 104–105
- public economics 155–157, 166–168, 189–191
- slavery 426–427
 - see also* elasticity of demand
- demand deposits 323–324, 328
- demand-pull model of inflation 338
- depreciation 76, 267, 288, 294–296, 539
- De Re Rustica* (Columella) 12–13, 15
- derived demand 41, 379–384, 426–427, 466
- deterioration 266–267, 287–289
- development 537–538
- differences of opinion 224–225
- differentiated goods 56, 79–82, 98
- direct taxation 151
- discount rates 273–275, 392
- discrete choice 77–78
- disinvestment 280–283
- distortions 154–155, 194
- diversification 230–235
- division of labor 535–536, 545–546
- dominance strategy 246
- durability of capital 265–266, 267
- durable goods
 - capital 263, 265–266, 267
 - consumption 56, 75–79, 97
 - housing 479
- Dutch auction 184–185
- dynamic equilibrium 264
- dynamic games 244–246
- dynamic labor supply 364–368
- dynamic optimization analysis 37
- earnings streams 355–356
- Easterlin model of fertility 557–558
- economic base theory 478
- economic growth *see* growth
- economic theories of government 195
- economies of scale *see* scale economies
- economies of scope 459
- edge of land use 443–445
- Edgeworth–Bowley box diagram 51–52, 127–128
- education 84, 218–223, 241–242
- efficiency
 - absolute and conditional efficiency 19
 - capital 266–267, 279
 - consumption 89
 - cost 43, 45
 - general equilibrium 127–128
 - growth 556
 - labor 366, 386–387, 402–405
 - natural resources 516
 - Pareto efficiency 140, 143, 145–146, 155
 - production 9, 11, 18–20
 - public economics 139–141, 143, 149–150, 154–155, 158, 172, 177
 - uncertainty 233
- efficiency wage hypothesis 386–387
- Egypt
 - consumption 72
 - foreign trade 117
 - industry structure 106, 117
 - labor 356, 423
 - production 21
- elasticity of demand
 - consumption 63–65, 79, 85, 93–94, 98
 - factors of production 41
 - general equilibrium 132–133
 - housing 481
 - labor 379–383, 401–403, 418
 - monopoly 108, 112
 - natural resources 523–524
 - public economics 157, 166–168, 170–171
- elasticity of substitution
 - consumption 60
 - general equilibrium 122–123
 - labor 361, 378–379
 - production 15, 16–18
- elasticity of supply 38–40
 - labor 360–361, 373, 380–382, 418
 - public economics 156
- embodiment 268
- endogenous centers 490–491
- endogenous growth 543–545
- endogenous money 335–336
- Engel curve 64–65
- English auction 184–185
- entry restriction 356–357
- epidemic disease 475
- equity 149–150
- equivalent variation 66–67
- error-learning model 246–249
- euergetism 173–174
- event risk 204, 209–210
- excess demand 120, 124–126, 134
- exchange 5, 302, 313–314, 335–337, 345
- excise tax 151–152, 153, 163
- excludability 141, 143–144
- exhaustability 141, 142
- exhaustible resources 516–524
- exogenous money 335–336
- expectations 246–252
 - adaptive models 246–249
 - capital 268, 285
 - cobweb model of agricultural supply 248, 250–251

- expectations (*continued*)
 - concepts and definitions 246–247
 - rational expectations hypothesis 230–233
 - resource-allocation decisions 247
 - risk 164, 207–215, 251
 - uncertainty 227
- expenditure function 418–419
- experts 223–224
- exploration 521–523
- exports *see* foreign trade
- extensive form representation 245–246
- externalities
 - cities 477
 - general equilibrium 134
 - natural resources 529
 - public economics 143–149
 - theory of 2
- factor intensity 129
- factor price 265
- factor shares *see* cost shares
- family
 - concepts and definitions 350–352, 398
 - family enterprise 414–423
 - fertility and child mortality 411, 412–414
 - implications of family enterprise 422–423
 - intrafamily resource allocation 293–294, 405–411
 - labor 350–352, 357–364, 368–375, 398–423
 - marriage 398–405, 428
 - missing markets and labor allocation 418–419
 - restrictions on household activities 420–422
 - separability of production and consumption decisions 415–418
 - supply 357–364, 368–375
- farmgate price *see* freight on board price
- federalism 196
- fertility 412–414, 557–558
- fiat system 262, 305, 335
- financial distribution 333
- financial intermediation 332–335
- firm-specific knowledge 353–354, 355
- first-order conditions
 - asymmetric information 238
 - cities 483–484, 490, 497
 - industry structure 105, 108, 110–112
 - labor 366, 370, 373, 411, 417–418
 - natural resources 518–519, 521–522
 - public economics 142–143, 156–157, 162, 174, 176
 - supply curves 35–36, 37
- Fisher, Irving 162–163
- fisheries 152–153, 218, 273–274, 463, 528–531
- fixed-coefficients technology 13, 17
- fixed costs 113
- fixed-point theorems 134
- fixed prices 90–93, 316–317, 322, 332, 342
- fixed quotas 420–422
- fixed supply 38
- flows
 - capital 263–264, 279–283
 - migration 396–398
 - money 301, 306
 - production 22–23
- foreign exchange 335–336, 345
- foreign trade
 - consumption 82
 - general equilibrium 121, 122, 135
 - industry structure 114–115, 117
 - public economics 169
- fourth moments 251–252
- free-rider problem 181, 185
- freight absorption 466
- freight on board price 69–70, 114–115, 443
- Friedman, Milton 317–318
- functional distribution of income 9, 23
- functional form 16–17
- Fundamental Theorem of Risk-Bearing 214, 227–228
- game theory 111, 185, 243–246
- general equilibrium 120–138
 - capital 283–284
 - cities 501
 - computable general equilibrium models 123, 134–136
 - concepts and definitions 120–124
 - disequilibrium 124
 - Edgeworth–Bowley box diagram 127–128
 - exchange 127–128
 - existence and uniqueness of equilibrium 133–134
 - fixed-point theorems 134
 - growth 548
 - growth in factor supplies 130–132
 - key facts 121
 - key questions 123–124
 - Lerner–Pearce diagram 128–133
 - multiple-sector models 122–123
 - one-sector model 122, 123–124
 - production 24–25
 - public economics 158, 166
 - real-world economies 123
 - stable and unstable equilibria 125–126
 - technical change 132–133
 - two-sector models 122, 123–124, 128–133
 - Walrasian model 122, 124–127
- geometric distributed lag 249
- Gini coefficient 23
- GNP *see* gross national product
- gold standard 307, 320–323, 335
- goods-in-process model 263
- Greece
 - labor 415, 423–424
 - land use 467–468
 - money 304, 319

- public economics 191–192
- religion 116
- gross complements 379, 382–383
- gross national product (GNP) 295–296
- gross substitutes 379, 382–383
- group decisions 224–225
- Groves–Ledyard–Clarke mechanism 185
- growth 535–567
 - ancient economies 3, 536, 544–545, 551, 553–559
 - capital 264, 283–284
 - capital-labor ratio 537–542, 556–560
 - concepts and definitions 535–538
 - delimiting the scope 535–536
 - development 537–538
 - economic growth theory 554–556
 - endogenizing technical change 543–545
 - evidence for 554–556, 559–561
 - following decline 557–559
 - governments 552–553
 - institutions 551–553
 - market division of labor and productivity 535–536, 545–546
 - natural resources 524–525
 - neoclassical growth theory 538–546
 - population models 557–559
 - production 536–546, 550–551, 555–557, 559–561
 - property rights 552
 - sectoral concepts as organizing devices 546–548
 - Solow model 538–544
 - stability and change 553
 - structural change 546–551, 561
 - technology 537, 541–545, 561
 - two-sector model of economy 548–549
- guilds 356–357
- Hamiltonian function 37
- Harberger model 166–167
- Harris–Todaro model 397–398
- harvesting natural resources 525–527
- health status 354, 356, 410–411
- hedonic price model 79–80, 93
- Henderson city-system model 493–503
- Hicksian demand curve 62, 67–68
- hierarchies of marketplaces *see* central place theory
- home workers 489–490
- horizontal equity 150
- horizontal integration 44–45
- hourly wages 386
- household production theory 5–6
 - consumption 78–79, 82–85
 - labor 350–352, 357–364, 369–375, 401–402, 415–423
- housing
 - bid-rent function 488–489
 - cities 473–474, 475, 479–488, 493–497, 506–507
 - demand 481–482, 494–496, 506–507
 - growth 555–556
 - housing market 19, 93–96, 475, 479–480, 482–488
 - public economics 152
 - spatialities 483–488
 - special characteristics of 479–480
 - supply 480–481, 486, 494–496, 506–507
- human capital
 - growth 543–544, 546, 556, 558–560
 - labor 351–352, 353–357, 393, 399, 409, 429
 - production 259–260
- hypothesis testing 222–223
- ideal theory of the state 552–553
- ideology 139
- imperfect markets *see* market imperfections
- imports *see* foreign trade
- impulsive behavior 88–89
- incentive compatibility constraint 239, 254
- incentives 4
- income
 - capital 264–265, 279–281
 - consumption 56, 63–65, 78–79, 83–85, 93–94
 - functional and personal distribution 9, 23
 - general equilibrium 120
 - growth 550–551
 - labor 358–368, 371–373, 400–405, 407–409, 413–414
 - money 321–322
 - neoclassical theory 23–24
 - production 8–9, 17–18, 23–25
 - public economics 158–161, 171–172
- income tax 152, 153, 158–161, 171–172
- income velocity 313–314
- incomplete markets 228
- indifference curve
 - capital 276–279
 - cities 484
 - consumption 58–61, 64, 68, 73–75, 77–78
 - general equilibrium 130–131
 - labor 357–362, 389, 408–409, 412–413, 416
 - public economics 159–160, 177–178
 - risk 213–215, 229
 - uncertainty 229
- indifference transitivity 183–184
- indirect taxation 151–152, 153
- indirect utility function 63
- individual equilibrium 212–215
- individual savings 294
- individual utility-family budget constraint model 364
- industry structure 103–119
 - competitive equilibrium 106–108
 - contestable markets 104, 112–113
 - markets 103
 - monopolistic competition 104, 111–112, 114–115
 - monopoly 103–104, 108–110, 115–117
 - monopsony 104, 113–114

- industry structure (*continued*)
 - oligopoly 104, 110–111, 115–117
 - perfect competition 103–106, 114–115
 - price 104
 - production 103
- industry supply curves 38
- inertia 88–89
- infant mortality 411, 414
- inflation 337–342
 - causes of 338–339
 - concepts and definitions 306–307, 337–338
 - consequences of 340–342
 - mechanisms of 339–340
 - neoclassical quantity theory 313–314
 - nominal and real distinctions 308–309
 - understanding of in antiquity 309
- information 217–225
 - asymmetric information 218, 235–242, 253–254
 - Bayesian probability theory 218–223
 - concepts and definitions 217
 - experts and groups 223–225
 - knowledge 217–218
 - learning 218–223
 - production 19, 25–26, 45
 - structure of 217–218
- inframarginal externalities 144–145
- infrastructure
 - capital 261
 - cities 493–497, 502, 504–505
 - production 21–22
 - public economics 188–191, 197
 - transportation 463–465
- innovation 535
- input demand curve 40–41
- input-output (i-o) model 122–123
- input price changes 20–21
- institutions 5, 551–553
- insurance 87–88
- intellectual property 218, 552
- interest rates
 - capital 272–284
 - consumption 76, 79
 - labor 365, 366
 - money 308–309, 313–314, 316, 328–331, 344
 - public economics 162–164, 187–188
- interindustry linkage 476–477
- internal rate of return (IRR) 188, 284–285
- intertemporal choice 56, 73–75
- intertemporal equilibrium 274–275
- intertemporal production prices 274
- intertemporal substitution effect 366
- intertemporal utility maximization 290–291, 527–528
- intrafamily resource allocation 293–294, 405–411
- inventories 260–261, 262–263
- inverse-elasticity rule 157, 177
- investment
 - capital 276–282, 284–287
 - labor 354–356, 427
 - slavery 427
 - theory of 285–287
- investment tax credit 166
- i-o *see* input-output model
- IRR *see* internal rate of return
- irrational behavior 55, 57, 88–89
- iso-profit curve 389–391
- isoquant diagram
 - cost 50–51
 - labor 376–378
 - production 13–15, 18–22
 - public economics 154–155
- joint probability 219–220
- judiciary systems 142
- Keynesian monetary theory 310, 315–317
- knowledge 217–218
- Koyk lag 249
- kurtosis 251–252
- labor 350–439
 - activity-related supply 368–369
 - allocation of time and resources 350–351, 352, 366–368, 371–373, 391
 - ancient economies 350–353, 392, 396–398, 415, 423–424
 - applying contemporary labor models 350–353
 - basis of pay 385–387
 - capital 259–260, 264–265, 269–270, 294, 537–542, 545–549, 556–560
 - children 407–414
 - cities 489–492, 494–495, 498–502
 - compensating differentials 350, 387–391
 - concepts and definitions 3, 350
 - consumption 69, 81, 84–85
 - contractual agreements 350, 352, 384–391
 - cost 29, 35, 40–43, 45–52
 - cost–benefit analysis 355, 392–393
 - demand 351–352, 375–384, 388–389, 401–403, 418–419, 426–427
 - derived demand 379–384, 426–427
 - economic theory of slavery 352, 354, 423–428
 - energy 383–384
 - family enterprise 414–423
 - family/household production 350–352, 357–364, 368–375, 398–423
 - fertility and child mortality 411, 412–414
 - general equilibrium 128–133
 - growth 537–542, 545–551, 556–560
 - guilds, licensing and entry restriction 356–357
 - health status 354, 356, 410–411
 - human capital 351–352, 353–357, 393, 399, 409, 429

- information problems and incentives 384–385
- intrafamily resource allocation 405–411
- investment in human capital 354–356, 427
- labor theory of value 269–270
- lifecycle/dynamic labor supply 364–368
- location theory 450, 454
- marriage 398–405, 428
- migration 352, 354, 391–398
- missing markets and labor allocation 418–419
- money 312–313
- multidimensional transactions 353
- natural resources 525–527, 529–530
- production 9, 11–13, 17–18, 20–23, 25–26
- public economics 158–161, 167, 171–172, 188
- separability of production and consumption
 - decisions 415–418
- sequencing of pay 387
- sleep and productivity 373–375
- supply 350–352, 357–375, 380–382, 388–389, 400–402, 416, 424–426
- uncertainty 234, 241–242, 252
- utility 352, 357–364, 366, 370–375, 405–410, 416–417
- Lagrange multiplier
 - behavioral uncertainty 235, 238–239
 - consumption 61–62
 - labor 365–366, 420–421
 - natural resources 518–519
 - public economics 170–171
 - supply curves 36–37
- land rent 443–445, 486–488
- land use 440–471
 - bid-rent function 447–450
 - central place theory 441, 455–456, 457, 461–463
 - consumption 62–63, 69
 - consumption and marketing location 441, 457–463
 - cost 29, 35, 41–43, 51–52
 - demand 441–442, 445–447, 460–461, 466
 - equilibrium in a region 450–451
 - general equilibrium 128–133
 - growth 547–551, 561
 - location theory 440–441, 442–468
 - manufacturing 441
 - modifying social context 451–452
 - periodic markets 462–463
 - production 9, 11–13, 17–18, 20–23, 25–26, 440–442, 452–457
 - public economics 153–154
 - special characteristics of land 440–441
 - supply 441, 445–447, 456–457
 - tenure 440
 - Thünen model 6, 26, 442–447, 488
 - transportation 441–444, 453–458, 463–467
 - see also* cities
- language 1, 2, 354
- large numbers problem 185
- Laspeyres index 70–71, 73
- law of variable proportions 11–13
- learning 84, 218–223, 241–242
- legislative politics 185–186
- leisure
 - cities 483–486
 - consumption 85
 - labor 352, 357–359, 361–362, 365, 418–419
 - public economics 159–160, 171
- lending 314
- Lerner–Pearce diagram 128–133
- Leviathan theory of government 195
- libraries 505
- licensing 356–357
- lifecycle labor supply 364–368
- lifecycle savings model 292–294
- likelihood matrix/function 220–222
- likelihood ratio 239
- Lindahl prices 143, 181–182, 186
- liquidity 302–303, 316, 323–324, 328–329
- liturgies 174
- localization economies 476
- local public goods 504–506
- location rent 443–445
- location theory 442–468
 - aggregate demand in spatial markets 460–461
 - bid-rent function 447–450
 - central place theory 441, 455–456, 457, 461–463
 - cities 482, 488
 - concepts and definitions 440–441
 - consumption and marketing location 441, 457–463
 - equilibrium in a region 450–451
 - individual facilities 452–454
 - industries 455–457
 - modifying social context 451–452
 - periodic markets 462–463
 - production facilities 452–457
 - shopping frequency versus storage 458–460
 - structure of transportation costs 457–458
 - Thünen model 442–447, 488
 - transportation 441–444, 453–458, 463–467
- logrolling 185
- long run
 - capital 272, 283
 - consumption 65
 - cost 29, 32–33
 - growth 545
 - labor 369, 376, 378–379
 - money 315
- loss offset provisions 164
- lump-sum distribution 140
- luxury products 24–25, 64
- macroeconomics 3, 311–313
- maintenance 287–289, 480–481
- Malthusian model of population 557–558

- manufacturing 441, 453–454, 550–551
- marginal benefit (MB) 142, 144–148, 187, 266
- marginal cost (MC) 31–33, 35
 - capital 266, 271
 - competition 108–109, 113
 - land use 456–457
 - natural resources 519, 527
 - public economics 142, 144–148, 158, 176, 178, 187
- marginal efficiency of investment 279
- marginal productivity 189
- marginal product (MP)
 - allocation across activities 43
 - growth 548
 - income 17–18, 24
 - labor 377–378, 401–403
 - substitution 11–12, 14
- marginal profit 34–35
- marginal rate of substitution (MRS)
 - consumption 59–60, 78
 - general equilibrium 127–128, 133
 - public economics 143, 154–155, 162, 174, 186, 193
 - risk 213
 - uncertainty 235
- marginal rate of technical substitution (MRTS) 14, 33, 34, 45, 154–155
- marginal rate of transformation (MRT) 127–128, 133, 154–155, 178–179, 193
- marginal reserve ratio (MRR) 331
- marginal resource cost (MRC) 114
- marginal revenue (MR) 35, 108–109, 114, 527
- marginal social cost (MSC) 147–149
- marginal value product 264
- markets 5
 - aggregate demand in spatial markets 460–461
 - asymmetric information 242
 - capital 269, 289–290
 - central place theory 441, 455–456, 457, 461–463
 - complete markets 227–228
 - consumption 56, 62, 77, 88–89, 441, 457–463
 - contestable markets 104, 112–113
 - contingent markets 225–230
 - general and partial equilibrium 120, 121
 - growth 535–536, 545–546, 556
 - income distribution 24
 - incomplete markets 228
 - industry structure 103–104, 112–113
 - labor 353, 370–373, 400–405, 414–415, 418–419, 427
 - land use 441–463
 - location of production facilities 452–457
 - location theory 442–463
 - market equilibrium 234–235
 - marketing location 441, 457–463
 - missing markets and labor allocation 418–419
 - money 315–316
 - natural resources 520–521, 527–528
 - periodic markets 462–463
 - production 19, 24
 - public economics 141–143, 180–181, 191–192
 - risk 204, 234–235
 - second-hand markets 77, 289–290
 - shopping frequency versus storage 458–460
 - slavery 427
 - structure of transportation costs 457–458
 - uncertainty 225–230, 242
 - see also* housing market
- market value 269
- marriage 398–405, 428
- Marshall–Hicks laws of derived demand 62, 68, 380–382, 426–427, 466
- Marxian theory of government 195
- material capital 259–262
- mathematical expressions 3
- matrix algebra 168–169
- maximum economic yield 529
- maximum sustainable yield 524–531
- MB *see* marginal benefit
- MC *see* marginal cost
- meaning of objects 96–97, 98
- mean-preserving spread 207–209
- mean-variance risk model 231
- Mesopotamia
 - capital 289
 - consumption 72
 - foreign trade 117
 - Isin workshop 44–45
 - labor 423–424
 - money 303–304
 - production 21
 - temple economy models 4
 - Ur pottery 43–44
- message services 217, 220–221
- microeconomics 2
- migration
 - allocation of time and resources 391
 - children 407–414
 - consequences of 394–396
 - economic incentives 392–394
 - Harris–Todaro model 397–398
 - labor 352, 354, 391–398
 - language 354
 - refugees 396
 - Roman Empire 392
- military equipment 465
- mining and mineralogy
 - capital 270
 - competition 117
 - labor 424
 - mining-forestry complex 531–532
 - money and banking 303, 318–323, 331–332, 338
 - natural resources 516–524, 531–532
 - public economics 153–154

- Minoan palaces 289
 Minoan pottery 30, 49–50
 minting 335–337, 343–345
 monarchy 553
 money 301–349
 ancient economies 309, 312–313
 bimetallism 337
 capital 262
 commodity money 304, 318–323
 concepts and definitions 301–302, 305–309
 contemporary synthesis 317–318
 creation by banks 323–328
 demand 301–302, 309–318, 321–322, 338–341, 345
 distinctiveness 311–312
 exchange 302, 313, 335–337, 345
 exogeneity and endogeneity of 335–336
 financial intermediation 332–335
 foreign exchange 335–336, 345
 inflation 306–307, 308–309, 313–314, 337–342
 interest rates 308–309, 313–314, 316, 328–331, 344
 Keynesian monetary theory 310, 315–317
 macroeconomics 311–313
 measurement of 309, 310–311
 monetary policy 342–345
 monetary theory 301–302, 312–313
 monetization prior to currency 303–304
 neoclassical quantity theory 310, 313–315
 neutrality of 314–315, 318
 nominal and real distinctions 307–309, 314
 price 305–309, 311, 314–316
 quantity equation 306, 313–314
 seigniorage 335–337, 343–345
 services of 302–304
 stability of value 303
 stocks and flows 301, 306, 310
 storing value 302–303, 315
 supply 301–302, 311–312, 318–341, 343–344
 types of 304–305
 unit-of-account 303
 see also banking
 monocentric city model 483–489
 monopolistic competition 80, 81–82, 104, 111–112, 114–115
 monopoly
 general equilibrium 134
 industry structure 103–104, 108–110, 115–117
 natural resources 523–524
 public economics 192–193
 monopsony 104, 113–114
 moral hazard 236, 237–239
 MP *see* marginal product; marginal productivity
 MRC *see* marginal resource cost
 MRR *see* marginal reserve ratio
 MRS *see* marginal rate of substitution
 MRT *see* marginal rate of transformation
 MRTS *see* marginal rate of technical substitution
 MSC *see* marginal social cost
 multiple-part pricing 178–179
 multiple-pass regeneration processes 10–11
 Mycenaean palatial system 4, 127, 289
 Mycenaean pottery 30, 47–49, 56–57, 96–99, 114–115
 Nash equilibrium 246
 national defense 142
Natural History (Pliny) 15
 natural resources 516–534
 ancient mining-forestry complex 531–532
 biological growth 524–525
 concepts and definitions 516–517
 different deposits 520–521
 exploration 521–523
 harvesting 525–527
 monopoly 523–524
 nonrenewable resources 516–524
 open access and the fishery 524, 528–531
 renewable resources 516, 524–531
 resource scarcity 531
 sustainability 524–531
 theory of optimal depletion 517–520
 theory of optimal use 527–528
 uncertainty 521
 neoclassical growth model 538–546
 endogenizing technical change 543–545
 market division of labor and productivity 535–536, 545–546
 Solow model 538–544
 technology and the Solow model 541–543
 neoclassical quantity theory 310, 313–315
 neoclassical theory of income distribution 23–24
 net present value (NPV) 187–188, 190–191, 284–285, 392
 neutrality of money 314–315, 318
 NMFIs *see* nonmonetary financial intermediaries
 noncooperation 406–407
 noncooperative game theory 111
 nonlabor income 361–364, 371–372
 nonmonetary financial intermediaries (NMFIs) 333–335
 nonrenewable resources 516–524
 nontradable goods 477–478, 493–494
 North, Douglass 552
 Noyelle–Stanback taxonomy 478–479, 503
 NPV *see* net present value
 obsolescence 288
 occupation-specific knowledge 353–354, 355
 oligopoly 104, 110–111, 115–117
 olive oil 82
 one hoss shay model 266
 one-sector general equilibrium model 122, 123–124

- open access resources 524, 528–531
- open city model 482–488
- operating costs 77
- opportunity cost 29, 42–43, 158
- optimal depletion, theory of 517–520
- optimal tax systems 157, 169–173, 185–186
- optimal use, theory of 527–528
- ordinary demand curve 62, 68
- overlapping generations model 293–294
- own-interest rates 273–275
- own-price elasticity of demand
 - consumption 63–65, 98
 - factors of production 41
 - labor 379, 382–383
 - public economics 157, 167–168, 170–171
- Paasche index 71, 73
- Pareto efficiency 140, 143, 145–146, 155, 174, 183–184
- partial elasticity of substitution 378–379
- partial equilibrium 24–25, 120–122, 155–158, 166, 283
- participation constraint 238–239
- p*-complementarity 382–383
- PDV *see* present discounted value
- pecuniary externality 149
- perfect competition 103–106, 114–115
- periodic markets 462–463
- permanent income hypothesis 265, 279–281, 291–292
- personal distribution of income 9, 23
- Phoenician littoral states 169
- piece rates 386
- poll tax 140
- Pompeii 93–96
- population models 369, 397, 557–559
- port authorities 464
- portfolio assets 165, 230–235, 334
- positive economics 185–186
- possibilities tradeoff curve 74
- posteriors 221–222
- pottery/ceramics
 - consumption 56–57, 96–99
 - cost 30, 43–44, 47–50
 - industry structure 114–115
 - natural resources 532
 - production function 10–11
- PPF *see* production possibilities frontier
- preferences
 - information 224
 - preference-revelation mechanisms 184–185
 - risk 211–213
 - transitivity 183–184
- preparedness maintenance 288–289
- present discounted value (PDV) 187–188, 190–191, 284–285, 392
- present-value relative prices 274–276
- preventive maintenance 288–289
- price 5–6
 - capital 264–265, 270–276, 286–287
 - cities 494–496
 - competitive behavior 252–253
 - competitive equilibrium 107–108
 - consumption 56–58, 60–63, 66–77, 83–84, 90–93
 - contestable markets 113
 - cost 35–36, 39, 43, 46–48, 52
 - expectations 246, 248–251
 - fixed prices 90–93
 - general equilibrium 125, 128–131, 135
 - industry structure 104, 107–113, 116–117
 - inflation 306–307, 308–309
 - labor 418–422, 425
 - land use 442–446, 459–461, 464–467
 - money 305–309, 311, 314–316, 321–322, 324
 - monopolistic competition 111–112
 - monopoly 109–110, 116–117
 - multiple-part pricing 178–179
 - natural resources 517–524, 527–528, 530–531
 - nominal and real distinctions 307–309
 - oligopoly 110–111, 116–117
 - price and consumption indexes 70–73
 - price theoretic models 5
 - production 20–21
 - public economics 142–143, 167–169, 175–181
 - Ramsey pricing model 176–178
 - risk 205, 214–215, 253
 - uncertainty 225–230, 252–253
- primal-dual relationship 29
- principal–agent relationships 195–196, 236, 237–239, 253–254, 384
- priors 221–222
- Prisoner's Dilemma 111, 185, 243–244
- private capital stock 261
- private goods 141–142, 173
- private munificence 173–174
- probability density function 206–208
- problem of the commons 218
- production 8–28
 - attributing products to inputs 17–18
 - capital 262–263, 274, 277, 284–285
 - cities 474, 475–479, 486, 493–497
 - competitive behavior 252–253
 - concepts and definitions 8
 - consumption 68–70, 78–79, 82–85
 - cost 29–30, 33, 36–52
 - demand for factors of 40–41
 - Edgeworth–Bowley box diagram 51–52
 - efficiency 9, 11, 18–20
 - entire economies 50–52
 - externalities 146–148, 477
 - functional form of production functions 16–17
 - general equilibrium 24–25, 129
 - growth 536–546, 550–551, 555–557, 559–561
 - income distribution 8–9, 17–18, 23–25

- industry structure 103, 106
- input price changes 20–21
- inputs and outputs 8, 16–21, 25
- isoquant diagram 13–15, 18–22
- labor 3, 352–354, 356, 369–379, 395, 399–402, 414–418
- land use 440–442, 452–457
- law of variable proportions 11–13
- measurement of substitution 15–16
- mode of production 18–20
- money 320–323, 329, 336–337
- multiple-pass regeneration processes 10–11
- organization of 43–46
- partial equilibrium 24–25
- policy 285
- predictions of 20–22
- production possibilities frontier 50–52
- public economics 139, 143, 146–148, 153–155, 168–169, 173, 175–191
- scale economies 475–477, 494
- stocks and flows 22–23
- substitution 8, 13–18
- surplus 68–70, 91–92
- technological changes 14, 21–22
- total product curve 11–12, 14
- types of 477–479
- uncertainty 228–229
- see also* household production theory
- production possibilities frontier (PPF) diagram 50–52
- product variety 56, 79–82, 98, 106
- profit maximization, cost 30, 34–36, 44
- profit-shares 386
- property rights 552
- property tax 153
- proportionality factor 234
- p*-substitutability 382–383
- public capital stock 261
- public economics 2, 139–201
 - Arrow impossibility theorem 183–184
 - Averch–Johnson effect 193–194
 - behavior of government and its agencies 194–197
 - bureaucracy 195–196
 - cities 495, 502–507
 - concepts and definitions 139–141
 - debt 173
 - externalities 143–149
 - government activities 139–141, 175–191
 - growth 552–553
 - levels of government 196
 - liturgies 174
 - local and national public goods 504–506
 - monetary policy 342–345
 - Pareto efficiency 140, 143, 145–146, 155, 174, 183–184
 - planning 179–181
 - positive economics of politics 185–186
 - positive/normative analysis 140
 - preference-revelation mechanisms 184–185
 - private goods 141–142, 173
 - private munificence 173–174
 - public goods 141–143, 175–186, 197, 398, 504–506
 - public investment and cost–benefit analysis 186–191
 - public production and pricing 175–181
 - raising revenue 139, 149–174, 196, 506–507
 - regulation of private economic activities 191–194
 - rent seeking 192–193
 - social choice mechanisms 181–185
 - structure of public enterprises 179
 - supply of public goods 181–186
 - taxation 139–141, 149–173, 182–183, 185–186
 - theories of government 194–195
 - theory of second best 174–175
 - Tiebout model 505–506
 - user fees 173
 - voting on public goods 183–184
- public goods 141–143, 175–186, 197, 398, 504–506
- pure efficiency effect 366
- q*-complementarity 382–383
- q*-substitutability 382–383
- quality/quantity tradeoffs 33, 82
- quantity theory *see* neoclassical quantity theory
- quasi-rents 270–272
- quotas 420–422
- Ramsey model of optimal taxation 157, 170
- Ramsey pricing model 176–178
- rates of return 164–165, 188, 230–233, 284–285
- Rathbone, Dominic 106
- rational expectations (RE) hypothesis 247, 249–252
- rationality 4, 55, 57, 88–89, 183–184
- Rawlsian social welfare function 150–151
- reaction curve 364
- recycling 454
- redistribution 171–173, 175, 177–178
- refugees 396
- regressive expectations model 249
- regulation 191–194
- relevance of economic theory 1–2, 4–6
- religion 115–117, 553
- remittances 395–396
- renewable resources 516, 524–531
- rent 41–43
 - asymmetric information 240
 - bid-rent function 447–450
 - capital 264–265, 270–272, 284
 - cities 486–487
 - consumption 62–63, 69, 75–76, 94
 - land use 443–451
 - public economics 191, 192–193
 - quasi-rents 270–272
 - rent seeking 192–193
- replacement costs 289–290, 428
- reputation 240

- reservation price 39
- reservation wage 359–362
- reserves
 - money and banking 304–305, 319–320, 323–332, 339, 344
 - natural resources 531
- revealed preference 57
- risk 202–217
 - Bayesian probability theory 216, 254
 - behavior 204, 209–215
 - capitalization of 234–235
 - certainty equivalence 210–213
 - competitive behavior 253
 - concepts and definitions 2, 202–203
 - consumption 60, 86–88
 - elements of risk 209–210
 - event risk 204, 209–210
 - expectations 251
 - expected utility 86–88, 210–215
 - expected value 207–209
 - indifference curve 213–215
 - individual equilibrium 212–215
 - labor 388–391
 - market risk 204
 - measurement of 205–209
 - natural resources 521
 - probability density function 206–208
 - public economics 164–165
 - risk aversion 86–88, 210–215, 388–389
 - risk premium 210–212
 - state-dependent consumption 213–214
 - state of nature 204, 209–210
 - substance of probabilities 215–217
 - ubiquity of risky decisions 203–204
 - uncertainty 202, 203–204, 210–217, 229–234
- Roman Empire
 - cities 489, 504–505
 - consumption 93–96
 - labor 392, 423–424
 - models of 4
 - money 319, 342
 - production 12–13, 19, 21–22
 - public economics 192
- Rotten Kid Theorem 400
- royalties 518–523, 527–529, 531
- Rybczynski theorem 131, 133
- salary 386
- sales quotas 420–422
- saving
 - capital 281–284, 290–294
 - growth 539, 542
 - individual and aggregate savings 294
 - intertemporal utility maximization 290–291
 - lifecycle savings model 292–294
 - permanent income hypothesis 291–292
 - public economics 161–164, 171–172
- scale economies
 - cities 475–477, 494
 - consumption 80–81
 - family 399
 - financial intermediaries 334
 - growth 556
 - land use 459
- scarcity value 520, 522, 529
- scrapping costs 289–290
- screening 236, 240, 242
- search models 253
- second best, theory of 174–175
- second-hand markets 77, 289–290
- second-order conditions 35–36, 37
- securities 333–335
- security 231
- seigniorage 335–337, 343–345
- separability 59
- shadow prices 62, 83, 188, 418–419, 422
- share wage payment 386
- Sharpe–Lintner capital asset pricing model 235
- shopping frequency 458–460
- short run
 - capital 272, 283
 - consumption 65
 - cost 29, 32–33
 - growth 545
 - labor 369, 376–378
- signaling 236, 240–242
- skewed distribution 23, 207, 208, 251–252
- skilled labor 85, 383–384, 393–394, 454
- slavery
 - ancient economies 423–424
 - capital 259–260
 - demand 426–427
 - domestic production and importing 425–426
 - economic theory of 352, 354, 423–428
 - investment 427
 - market consequences 427
 - money 312
 - slaves' incentives 427–428
 - supply 424–426
- sleep 373–375
- Slutsky equation 64, 158, 162, 170, 172
- small open economy 169
- social choice mechanisms 181–185
- social welfare function 150–151
- Solow model 538–544
- spatialities
 - cities 473, 482–492, 498–499
 - closed city model 482, 487–488
 - density gradients in ancient cities 491
 - endogenous centers 490–491
 - home workers 489–490
 - monocentric city model 483–489
 - multiple resident categories 488–489
 - open city model 482–488

- wage differentials across cities 491–492, 498–499, 501–502
- specialization 556
- specificity 5
- speculative demand 315
- Stackelberg behavior 111
- standard deviation 208, 232–233
- state-claims 214–215, 226–230
- state-dependent consumption/utility 213–214, 229–230
- static equilibrium 263–264, 281–282
- static games 243–244
- static optimization analysis 37
- stock markets 315–316
- stocks 263–264, 279–283
 - money 301, 310
 - natural resources 517–531
 - production 22–23
- Stolper, Matthew 25–26
- storage 458–460
- strategic behavior 242–246
- substitutes in production 379–380, 382–383
- substitution 5
 - consumption 59–60, 64, 78, 80, 84–85, 96–98
 - cost 33, 34, 45
 - functional form of production functions 16–17
 - general equilibrium 122–123, 127–128
 - income distribution 18
 - labor 360–361, 366–367, 372–374, 378–384
 - land use 447, 448, 453–454
 - measurement of 15–16
 - natural resources 519–520
 - production 8, 13–18
 - public economics 143, 159, 162–167, 174, 186
 - risk 213
 - uncertainty 235
 - see also* elasticity of substitution
- sunk costs 113
- supply
 - capital 279–283
 - competitive equilibrium 106–108
 - consumption 90–91
 - cost 35–40, 42
 - expectations 248–253
 - general equilibrium 121, 125–126
 - housing 480–481, 486, 494–496, 506–507
 - labor 350–352, 357–375, 380–382, 388–389, 400–402, 416, 424–426
 - land use 441, 445–447, 456–457
 - money 301–302, 311–312, 318–341, 343–344
 - monopsony 114
 - natural resources 521
 - public economics 147, 156, 181–186
 - slavery 424–426
 - supply/demand shifters 107–108
- surplus 67–70, 91–92, 127, 190
- sustainability 524–531
- systems of cities 492–503
 - city size distribution 499–503
 - concepts and definitions 492–493
 - different types of cities 497–499
 - production and consumption 493–497
- tabernae* 95
- tastes 57
- taxation
 - acceptability 153
 - categories 151–152
 - cities 506–507
 - deadweight losses 155–158
 - distortions 154–155
 - effects of taxes 154–165
 - equity versus efficiency 149–150
 - general equilibrium 121, 124
 - implementation and administration 152
 - income tax and labor supply 158–161
 - money 341
 - optimal provision of public goods 182–183
 - optimal tax systems 157, 169–173, 185–186
 - public economics 139–141, 149–173, 182–183, 185–186
 - rationales and instruments 149–154
 - redistribution 171–173
 - risk taking 164–165
 - saving 161–164, 171–172
 - social welfare function 150–151
 - tax incidence and liability 165–169
- TC *see* total cost
- technology
 - capital 268, 288
 - consumption 81, 85
 - cost 33, 36–38, 40–41, 45–46
 - general equilibrium 132–133
 - growth 537, 541–545, 561
 - labor 383, 395
 - land use 442–443
 - natural resources 516, 519–520, 529–531
 - production 8, 10, 13–15, 19, 21–22
- temporary equilibrium 281–282
- tenure 440
- theory of capital 276–284
- theory of investment 285–287
- theory of optimal depletion 517–520
- theory of optimal use 527–528
- theory of second best 174–175
- third moments 251
- Thünen model of land use 6, 26, 442–447, 488
- Tiebout model 505–506
- tied quotas 420–422
- time allocation *see* household production theory
- total cost (TC) curve 31–32
- total product (TP) curve 11–12, 14
- total variable cost (TVC) 31–32
- TP *see* total product

- training 353–356
- transcendental logarithmic (translog) function 17
- transformation curve 229, 277, 409
- translog *see* transcendental logarithmic
- transportation
 - cities 482
 - consumption 69–70
 - equipment 465
 - industry structure 114–115
 - infrastructure 463–465
 - location theory 441–444, 453–458, 463–467
 - natural resources 532
 - pricing of transportation services 465–467
 - structure of transportation costs 457–458
- TVC *see* total variable cost
- two-period model of consumption 162–164, 171–172
- two-sector general equilibrium models 122, 123–124, 128–133
- uncertainty
 - asymmetric information 235–242
 - behavioral uncertainty 235–246
 - capital 288–289
 - competitive behavior 252–253
 - complete markets 227–228
 - consumption 56, 79
 - contingent markets 225–230
 - dealing with nature's uncertainty 225–235
 - incomplete markets 228
 - labor 356
 - market equilibrium and capitalization of
 - risk 234–235
 - maximizing state-contingent utility 226–227
 - money 328–329
 - natural resources 521
 - portfolios and diversification 230–235
 - production adjustments for contingent
 - consumption 228–229
 - risk 202, 203–204, 210–217, 229–234
 - state-claims and assets 226–230
 - state-dependent utility 229–230
 - strategic behavior 242–246
- unconditional probability 219, 221
- uniform commodity tax 151–152, 153, 163
- unions 357
- unskilled labor 383–384
- Ur 319
- urban economies *see* cities
- user cost 76–77
- user fees 173, 506
- utility
 - asymmetric information 238–239
 - capital 290–291
 - cities 483–492, 496–497
 - consumption 58–60, 61–63, 66–69, 77–88
 - expected utility 86–88
 - intertemporal utility maximization 290–291
 - labor 352, 357–364, 366, 370–375, 405–410, 416–417
 - money 311, 317
 - public economics 142–145, 147–148, 163–165, 171–172, 174
 - risk 209–215
 - strategic behavior 246
 - uncertainty 226–230, 236, 238–239
- value
 - capital 264, 269–270, 280–282
 - labor theory of value 269–270
 - land use 451
 - money 302–303, 307–308, 315
 - nominal and real distinctions 307–308
- value of marginal product (VMP) 43, 377–378, 387, 441–442
- Van de Moortel, Aleydis 49–50
- Van Wijngaarden, Gert-Jan 47–49, 96–99
- variable factor 12
- variable proportions, law of 11–13
- variance 208, 231
- vertical equity 150
- vertical integration 45
- Vickrey auction 185
- viticulture 12–13
- VMP *see* value of marginal product
- von Neumann–Morgenstern utility function 86, 209, 210
- voting theory 183–184
- wages 41–43
 - asymmetric information 241–242
 - capital 264–265
 - concepts and definitions 350
 - consumption 56, 83–85
 - contractual agreements 384–391
 - differentials across cities 491–492, 498–499, 501–502
 - family 400–405, 407–409, 413–414
 - migration 392–398
 - supply 359–368, 371–373
- Wallace-Hadrill, Andrew 93–95
- Walrasian model 122, 124–127
- Walras' Law 125, 321–322
- water 442, 452, 454, 504–505
- wealth taxes 153
- weapons 260, 261–262
- welfare state 140, 150–151, 170
- wellbeing 55, 59, 66–70
- withdrawal rates 324–325, 329
- Z-goods 370–375, 401–402